FUNDAMENTAL CONCEPTS OF ALGEBRA

by
BRUCE E. MESERVE

DOVER PUBLICATIONS, INC., NEW YORK

Published in Canada by General Publishing Company, Ltd., 30 Lesmill Road, Don Mills, Toronto, Ontario.
Published in the United Kingdom by Constable and Company, Ltd., 10 Orange Street, London WC2H 7EG.

This Dover edition, first published in 1982, is an unabridged and slightly corrected republication of the second printing (1959) of the work originally published by Addison-Wesley Publishing Company, Inc., Reading, Mass., in 1953.

Manufactured in the United States of America
Dover Publications, Inc., 180 Varick Street, New York, N.Y. 10014

Library of Congress Cataloging in Publication Data

Meserve, Bruce Elwyn, 1917-
Fundamental concepts of algebra.

Reprint. Originally published: Cambridge, Mass. : Addison-Wesley, 1953 (1959 printing) With slight corrections.
Bibliography: p.
Includes index.
1. Algebra. I. Title.
QA154.2.M47 1982 512 82-9411
ISBN 0-486-61470-0 (pbk.) AACR2

PREFACE

This book is based upon the algebra section of a course entitled the "Fundamental Concepts of Mathematics" as it has evolved at the University of Illinois during the past forty years. Topics from analysis are developed and used along with those from the algebra. The part of the above course that is not included here is primarily concerned with the fundamental concepts of geometry and is offered as an independent unit.

This book adopts a modern viewpoint of the algebra and the analysis. It recognizes a basic need for a knowledge of the fundamental concepts of these subjects apart from what is gained in the specialized courses in each of their many subdivisions. Such a need is especially felt by prospective teachers of secondary mathematics, students preparing for specialized advanced undergraduate courses in mathematics, and anyone desiring a broad liberal education.

The author and others have used this book in mimeographed form as a text at the advanced undergraduate-graduate level for several years. Most of the students have previously studied college mathematics through the calculus. However, this book has also been successfully used where the calculus was not a prerequisite. There is ample material for a course having 45 class hours.

The development of the complex number system and the elementary theories of numbers, polynomials, and equations (Chapters 1–4) are discussed using the concepts and terminology of modern algebra. Chapters 5 (Determinants and Matrices) and 6 (Constructions) depend upon the first four chapters but are independent of each other. The broad scope of this book is made possible by considering primarily the fundamental concepts of the various subjects. These concepts are illustrated by numerous examples, and the theory is frequently extended by suitable sequences of exercises. Basically this book is concerned with the fundamental concepts of higher mathematics (algebra and analysis) with relation to elementary mathematics. In this way it aims to introduce the concepts of higher mathematics and bring the reader to a thorough understanding of elementary mathematics.

The author is deeply indebted to Professor E. B. Lytle, Professor Echo Pepper, and Professor J. H. Chanler for their part in the

evolution of the course on which this book is based. Professor Chanler has in addition used this book in mimeographed form as a text in her classes and offered many valuable suggestions throughout the preparation of the manuscript. Acknowledgments are also due to the constructive criticisms of many students, to Professor F. E. Hohn who read the manuscript, to my wife who typed the manuscript, and to the publishers for their cooperation and very efficient service. To each and all the author is sincerely grateful.

B. E. M.

December, 1952

CONTENTS

Starred sections may be omitted without disturbing the organization of the text.

CHAPTER 1.	OUR NUMBER SYSTEM	1
1–1	Sets	1
1–2	Cardinal numbers	3
1–3	Equivalence relations	7
1–4	Peano's postulates	8
1–5	Addition and multiplication	10
1–6	Order relations	14
1–7	Inverses	15
1–8	Positive rational numbers	17
1–9	Negative numbers	20
1–10	Real numbers	24
*1–11	Postulates for real numbers	28
1–12	Properties of real numbers	30
1–13	Transfinite cardinal numbers	35
1–14	Group; number system	39
1–15	Complex numbers	40
1–16	Properties of complex numbers	45
1–17	De Moivre's Theorem	50
*1–18	Fields and number systems	54
CHAPTER 2.	THEORY OF NUMBERS	58
2–1	Divisibility	58
2–2	Division Algorithm	60
2–3	Prime numbers	63
2–4	Unique Factorization Theorem	67
2–5	Euclidean Algorithm	71
2–6	Bases	75
2–7	Decimal notation	80
*2–8	Congruences	83
*2–9	Residue classes. Euler's ϕ-function	87
*2–10	Evaluation of $\phi(m)$	90
*2–11	Linear congruences	93
*2–12	Diophantine problems	95

CHAPTER 3. THEORY OF POLYNOMIALS 98
3–1 Polynomials 98
3–2 Rings of polynomials 100
3–3 Rational functions 101
3–4 Divisibility 102
3–5 Division Algorithm 104
3–6 Irreducible polynomials 106
3–7 Euclidean Algorithm 109
3–8 Change of variable 112
*3–9 Ideals 113
3–10 Functions 114
3–11 Limits 118
3–12 Continuity 122
3–13 Continuous functions 124
3–14 Derivatives 127
*3–15 Taylor's series 131
*3–16 Analytic functions 132

CHAPTER 4. THEORY OF EQUATIONS 134
4–1 Zeros of a polynomial 134
4–2 Synthetic division 135
4–3 Change of variable 138
4–4 Number of roots 139
4–5 Determination of the roots 141
4–6 Conjugate imaginary roots 142
4–7 Elementary symmetric polynomials 144
4–8 Transformations of roots 146
*4–9 Cubic equations 150
*4–10 Quartic equations 154
4–11 Descartes' Rule of Signs 156
4–12 Sturm's Theorem 160
4–13 Multiple roots 164
4–14 Approximate solutions 169

CHAPTER 5. DETERMINANTS AND MATRICES 172
5–1 Historical development 172
5–2 Matrices 173
5–3 Permutations 175
5–4 Inversions 177
5–5 Transpositions 178

5–6 Even and odd permutations 181
5–7 Determinants 183
5–8 Properties of determinants 187
5–9 Expansion of determinants 192
5–10 Minors 198
5–11 Cramer's Rule 205
5–12 Systems of linear equations 208
5–13 Linear dependence 212
5–14 Applications in analytic geometry 217
5–15 Geometric transformations 221

CHAPTER 6. CONSTRUCTIONS 228
6–1 Classical constructions 229
6–2 Elementary classical constructions 231
6–3 The algebraic viewpoint 232
6–4 Basic classical constructions 233
6–5 Construction of roots of equations 236
6–6 Famous construction problems 238
6–7 Nonclassical geometric trisections 242
6–8 Mechanical angle trisectors 244
6–9 Linkages 246
6–10 Summary 247

CHAPTER 7. GRAPHICAL REPRESENTATIONS 250
7–1 Euclidean and complex spaces 250
7–2 Polynomials 252
7–3 Conic sections 254
7–4 Quadric surfaces 258
7–5 Higher plane curves 262
7–6 Rational functions 264
7–7 Algebraic functions 268
7–8 Curve tracing 270
7–9 Special graphs 274
7–10 Graphical solutions 276
7–11 Curve fitting 277
7–12 Conclusion 280

BIBLIOGRAPHY 281

SYMBOLS AND NOTATION 285

INDEX 287

CHAPTER 1

OUR NUMBER SYSTEM

Nearly everyone uses a number system daily, yet few people can describe a number system accurately. We shall endeavor in this chapter to gain an appreciation of numbers and some of the relations among them. The following is a presentation of one method of developing the three number systems that are most commonly used today: the rational, the real, and the complex number systems. It will become evident in this development that these three systems are related in that the real numbers include the rational, and the complex include the real and the rational. A few other number systems will be mentioned briefly.

1–1 Sets. Numbers are frequently associated with sets of objects. Three men, three stones, three logs have in common a property that may have been first indicated by ///. We shall introduce a few properties of numbers in terms of sets (defined below) and correspondences between sets. Later we shall adopt a postulational approach as a basis for a more intensive study of numbers.

The concept of a set, class, aggregate, . . . of elements is fundamental, not only in mathematics but also in daily living. For example, one often considers a pair of shoes, a set of golf clubs, a set of chessmen, a deck of cards, a set of books, a set of tires for a car, etc. In mathematics one might consider the set of positive integers, the three vertices of a triangle, the set of roots of a polynomial equation, the set of positive even integers less than 1000, the set of positive prime numbers, the totality of real numbers, etc. Formally, we shall paraphrase G. Cantor and define a *set** S as a collection into a whole of distinct, perceived, or considered objects called the *elements* of S. In practice, the reader's grasp of the full meaning and importance of this concept (as well as many others) will develop as extensive use is made of it.

---

* Throughout this text new terms will be italicized when they are defined or first identified.

The numbers that primitive man first used in counting the elements of a set of objects are called *natural numbers* or *positive integers.* Technically, the positive integers are symbols. They may be written as /, //, ///, . . . ; i, ii, iii, . . . ; 1, 2, 3, . . . ; or in many other ways. There also exist many other symbols, such as 0, -3, $\sqrt{2}$, and π, which we shall later define to be numbers, i.e., we shall extend the meaning of "number" to include symbols that are not positive integers. First, however, we shall consider some of the basic properties of positive integers.

When the positive integers are used to count the elements of a set, they are sometimes called *ordinal numbers;* when they are used to designate the number of elements in a set, they are called *cardinal numbers.* We shall consider the concept of a cardinal number in terms of the common properties of the sets in any class of sets that have the same cardinal number.

The common property of the sets of three men, three stones, three logs may have been first observed when each man had a stone in his hand or sat on a log. This common property is most easily understood in terms of one-to-one correspondences, another fundamental concept of mathematics. There is a *one-to-one correspondence* between the elements of two sets A, B whenever each element of the set A corresponds to exactly one element of the set B, and each element of B is the correspondent of exactly one element of A. The cardinal number b of any set B designates a common property of all sets A_a such that the elements of each set A_a may be placed in one-to-one correspondence with the elements of B.

We may associate the same number, 3, with each of the sets of men, stones, and logs if and only if we may count the elements of each set using the integers 1, 2, 3, i.e., if there is a one-to-one correspondence between each set and the set of positive integers 1, 2, 3. Thus the number 3 represents a common property of the sets of three integers, three men, three stones, three logs, and any other set of elements that can be placed in one-to-one correspondence with any one of these sets. In other words, all sets that can be placed in one-to-one correspondence with the set 1, 2, 3 have a common property that is designated by the cardinal number 3. In this sense the cardinal number 3 denotes an arbitrary set of this class of sets. This concept and that of one-to-one correspondences between sets form the basis for our discussion of properties of cardinal numbers.

Exercises

1. Give or describe two sets of elements having cardinal number 4 and indicate how a one-to-one correspondence may be obtained between them.
2. Repeat Exercise 1 using a different pair of sets having cardinal number 4.
3. Repeat Exercise 1 for the cardinal number 10.
4. Repeat Exercise 1 for the cardinal number 20.

1-2 Cardinal numbers. If for a given set of elements S there exists a positive integer N such that the elements of S may be placed in one-to-one correspondence with the set of positive integers $1, 2, \ldots, N$, we say that S is a *finite set* with (finite) cardinal number N. If there does not exist a positive integer N with this property, and if S has at least one element, we say that S is an *infinite set.* The cardinal number of any finite set may be obtained by counting the elements of the set, i.e., it corresponds to the greatest ordinal number used in counting the elements of the set. The concept of a cardinal number as an arbitrary representative of a class of sets makes it possible to associate transfinite cardinal numbers with infinite sets (Section 1-13).

Comparisons between cardinal numbers must agree with the corresponding comparisons between the sets of elements represented by the cardinal numbers. Accordingly, the cardinal numbers a, b associated with sets A, B are equal (written $a = b$) and the sets are said to be *equivalent* if there exists a one-to-one correspondence between the elements of the two sets. The cardinal number a is *less than* the cardinal number b (written $a < b$) and b is *greater than* a (written $b > a$) if after associating each element of A with an element of B (one-to-one) there remains at least one element of B that has not been associated with an element of A, and there does not exist a one-to-one correspondence between the elements of B and the elements of A. The second condition is superfluous in the case of finite sets but necessary for infinite sets. For example, if both of the sets A, B consist of the set of all positive integers n, there exists the one-to-one correspondence (n to n) of each integer with itself, and the set A has the same cardinal number as the set B. However, there also exists the one-to-one correspondence (n to $2n$) of all the integers in A with the even integers in B. In this correspondence between the infinite sets there remain elements of B (the odd integers) that have not been associated with elements of A. We shall consider this problem in more detail in our discussion of transfinite cardinal

numbers (Section 1–13). For an example in the case of finite sets, let A be the set of students in a class and B the set of chairs in the classroom. If each student has a chair and each chair is occupied by a student, then $a = b$. If each student has a chair and at least one chair is unoccupied, then $a < b$. If each chair is occupied and at least one student does not have a chair, then $a > b$.

A set of elements B is called a *subset* of a set A if each element of B is an element of A, a *proper subset* if it is a subset and there is at least one element of A that is not an element of B. The set that does not contain any element is called the *null set* or *empty set* and is considered a subset of every set. Using this terminology, $a = b$ if A is equivalent to a subset of B and B is equivalent to a subset of A; $a < b$ if A is equivalent to a proper subset of B and B is not equivalent to any subset of A. Given any two finite sets A, B with cardinal numbers a, b, we may compare the cardinal numbers using the subsets $1, 2, \ldots, a$ and $1, 2, \ldots, b$ of the set of positive integers. Let C be the set $1, 2, \ldots, c$ of positive integers that are in both these subsets. If $c = a$ and $c \neq b$, then $a < b$. If $c = a$ and $c = b$, then $a = b$. If $c = b$ and $c \neq a$, then $b < a$. Thus we have proved that for any two finite sets A, B with cardinal numbers a, b exactly one of the relations $a < b$, $a = b$, $a > b$ must hold.

The above example of students may be extended to illustrate the addition of cardinal numbers. Let G be the set of girls in the class, B the set of boys, C the set of chairs, and g, b, c the respective cardinal numbers of these sets. If each student has a chair and each chair is occupied by a student, then $c = g + b$. In general, given sets A, B, C, where A and B have no elements in common (i.e., the sets A and B are *mutually exclusive*), we write $a + b = c$ when there is a one-to-one correspondence between the elements of C and the totality of elements of A and B; i.e., C is equivalent to $A + B$ where addition of sets is understood to be in a *set-theoretic* (totality of elements) sense. Thus the addition of any two cardinal numbers may be easily understood in terms of one-to-one correspondences. Multiplication may also be defined for cardinal numbers. Subtraction and division may be defined only in special cases. For example, we may write $c - b = a$ if and only if there exists a cardinal number a such that $c = a + b$.

The product of two cardinal numbers, like the product of two positive integers, may be expressed using the concept of successive addition: $1 \cdot a = a$, $2 \cdot a = a + a$, $3 \cdot a = a + a + a, \ldots$. If the

number of boys is equal to the number of girls in the class discussed above, then $g = b$ and $c = b + b = 2 \cdot b$. In this case, the product $ab = 2b$ is the cardinal number of a set C equivalent to the set-theoretic sum of the set C_1 of chairs occupied by the girls and the set C_2 of chairs occupied by the boys. In general, we write $c = ab$ whenever C is equivalent to a set-theoretic sum of mutually exclusive sets $C_1, C_2, \ldots, C_a$; each C_i is equivalent to B, and there exists a one-to-one correspondence (which we have indicated by the subscripts) between the elements of A and the set of sets C_i. In the above example C is equivalent to $B + G$, B is equivalent to G, and there exists a one-to-one correspondence between the elements of A, say a_1, a_2, and the set consisting of the two elements B, G. We may also write $c/b = a$ whenever $c = ab$.

The above four *rational operations* (addition, subtraction, multiplication, division) will be considered extensively throughout this text. In the case of cardinal numbers we have seen that the sum of any two cardinal numbers is a cardinal number, the difference between two cardinal numbers is a cardinal number whenever it is defined, the product of any two cardinal numbers is a cardinal number, and the quotient of two cardinal numbers is a cardinal number whenever it is defined. Furthermore, our definitions are sufficient to enable us to prove (a) that cardinal numbers satisfy the usual order relations for positive integers (Exercises 7 and 8), and (b) that addition (Exercise 9) and multiplication (Exercise 10) of cardinal numbers have the basic properties that we shall expect for addition and multiplication of positive integers (Section 1–5).

Before discussing the properties of positive integers it appears desirable to discuss briefly the words "operation" and "relation." Given any two elements a, b of a set S, we often associate with them other elements such as $a + b$, $a - b$, $a \cdot b$, a/b of S. Such operations are called *binary operations* in S. In general, a set S is *closed* under a binary operation $\oplus$, and the operation is *uniquely defined* over the set S if for all elements a, b in S the element $a \oplus b$ is a unique element of S. The binary operations of addition and multiplication have been defined over the set of cardinal numbers.

We may also compare two elements a, b of a set S. For example, $a > b$ and $a = b$ indicate comparisons or *binary relations* among elements of S. A binary relation $\ominus$ is defined over a set S if for every ordered pair (a, b) of elements of S it can be determined whether or not the relation holds. We shall assume that any binary

relation must either hold or not hold. Basically, we shall assume that given any two numbers a, b, exactly one of the relations $a = b$, $a \neq b$ must hold. Throughout this text we shall be concerned with binary operations and binary relations. In general, the set S will be specified. The set S might be a particular set of numbers or a set of polynomials in certain specified variables with coefficients from a particular set of numbers. We shall endeavor to specify or characterize each relation used in terms of its basic properties, i.e., we shall state properties of the relation such that all statements involving the relation will hold for all relations having these properties.

The binary operations of addition and multiplication will be treated in the manner described above, i.e., we shall endeavor to characterize these operations by means of their basic properties. Accordingly, the development of our number system considered in this chapter is essentially a consideration of the basic properties of equivalence relations, positive integers, addition, multiplication, order relations, inverse numbers, inverse operations, positive rational numbers, negative numbers, real numbers, and complex numbers. The order of the topics in this development follows closely that in the historical development of our number system. The postulational approach represents a comparatively recent mathematical formalization of the subject that emphasizes the fundamental concepts upon which algebra is based [10; 221–232].* The approach from the theory of sets is also comparatively recent. It is discussed in [28]. A nontechnical discussion of the development of the number concept with many historical anecdotes may be found in [17].

EXERCISES

1. Use sets of elements to give an example of the addition of cardinal numbers.

2. Use sets of elements to give an example of the subtraction of cardinal numbers.

3. Is $a - b$ defined for all cardinal numbers? Explain.

4. Give an example of sets A, B satisfying each of the following: (a) $a = b$, (b) $a = 2b$, (c) $a = 4b$, (d) $a < b$.

5. Give an example of sets A, B satisfying each of the following: (a) $a - b$ is defined, (b) $a - b$ is not defined, (c) a/b is defined, (d) a/b is not defined.

* The symbol [10; 221–232] is used to refer to pages 221–232 of reference number 10 in the list of references at the end of this book.

6. Using your knowledge of the integers, indicate a one-to-one correspondence between the positive integers and the (a) positive even integers, (b) negative integers, (c) positive integral multiples of ten, (d) positive integral powers of two.

7. Prove that for any cardinal numbers a, b, c (a) $a < b$ and $b < c$ imply $a < c$, (b) $a < b$ implies $a + c < b + c$, (c) $a < b$ implies $ac < bc$.

8. Define $a \geqq b$ for arbitrary cardinal numbers a, b and repeat Exercise 7 for the relation $\geqq$.

9. Prove that for any cardinal numbers a, b, c (a) $a + b$ is a unique cardinal number, (b) $a + b = b + a$, (c) $(a + b) + c = a + (b + c)$.

10. Prove that for any cardinal numbers a, b, c (a) ab is a unique cardinal number, (b) $ab = ba$, (c) $(ab)c = a(bc)$, (d) $(a + b)c = ac + bc$.

1–3 Equivalence relations. Any relation having the three properties:

reflexive, $a = a$,
symmetric, $a = b$ implies $b = a$,
transitive, $a = b$ and $b = c$ imply $a = c$,

is called an *equivalence relation*. The equivalence of sets and therefore the equality of cardinal numbers as defined in Section 1–2 can be proved to be an equivalence relation as follows. It is reflexive, since the elements of any set may be placed in one-to-one correspondence with themselves. It is symmetric, since any one-to-one correspondence between the elements of a set A and the elements of a set B may also be considered as a one-to-one correspondence between the elements of the set B and those of the set A. Finally, it is transitive, since a one-to-one correspondence between the elements of a set A and those of a set B and a second one-to-one correspondence between the elements of B and those of a set C give rise to a one-to-one correspondence between the elements of A and those of C. For example, in the case of finite sets, if we designate the corresponding elements of A, B, C by a_i, b_i, c_i respectively, we obtain correspondences similar to those indicated in Fig. 1–1.

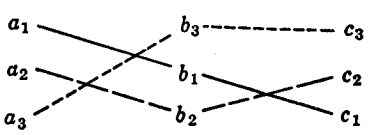


FIG. 1–1

One can also prove under the usual definitions that "identity" ($\equiv$), "congruence" ($\cong$) of geometric figures, and "similarity" ($\sim$) of geometric figures are equivalence relations. Thus each of

the symbols $=$, $\equiv$, $\cong$, $\sim$ represents "equals" in a well-defined mathematical sense. We now use the equivalence relation $=$ in a characterization of the positive integers by means of Peano's postulates. As indicated in Section 1–2, we assume that given any two numbers a, b exactly one of the relations $a = b$, $a \neq b$ must hold.

EXERCISES

1. Is $<$ an equivalence relation? Explain.
2. Prove that similarity of figures in plane geometry is an equivalence relation.
3. Is $\neq$ an equivalence relation? Explain.
4. State which of the following relations among students are equivalence relations: (a) being of the same age, e.g., Ruth is the same age as John, (b) being older, (c) being at least as old as, (d) being of the same weight, (e) being of different weights, (f) making better grades, (g) having any given characteristic in common, (h) not having a given characteristic in common.
5. Is the property of being different an equivalence relation among people? Explain.
6. Give an example of a relation that is transitive but neither reflexive nor symmetric.
7. Give an example of a relation that is reflexive and transitive but not symmetric.
8. Give an example of a relation that is symmetric but neither reflexive nor transitive.
9. Prove that, considered as a set of elements, the set of positive integers is equivalent to the set of positive integral powers of ten.
10. Illustrate the equivalence between the set of points on a line segment one unit long and the set of points on a line segment ten units long.
11. Illustrate the equivalence between the set of points on a line and the set of points on a circle with one point removed.
12. Illustrate the equivalence between the set of points on a plane and the set of points on a sphere with one point removed.

1–4 Peano's postulates. We now start our discussion of a logical development of our number system. In this section we shall first state five properties that may be used to characterize the positive integers, then assume that the positive integers have these properties (i.e., take these properties as postulates for the development of our number system), and finally use one of the postulates to obtain a formal procedure for proving that a relation holds for all positive integers. The following five statements are known as *Peano's postulates:*

(i) 1 is a positive integer.

(ii) Every positive integer a has a unique positive integer a^+ as its successor.

(iii) No positive integer has 1 as its successor.

(iv) If $a^+ = b^+$, then $a = b$.

(v) Every set of positive integers that contains 1 and the successor of every positive integer in the set contains all positive integers.

Postulate (v) is sometimes called the *principle of complete induction.* It provides a basis for the principle of mathematical induction. Since every positive integer a has a successor a^+, there is no largest positive integer and it is not possible to verify any relation for each and every positive integer separately. Accordingly, we need to use the principle of complete induction to prove that a relation or proposition is valid for all positive integers n. In particular, we consider the set S of positive integers for which the proposition holds (is valid). If 1 is in the set S and, for each positive integer k in S, the integer $k^+ = k + 1$ is also in S, then by the principle of complete induction all positive integers are in the set S, i.e., the proposition is valid for all positive integers. Formally, we have the *principle of mathematical induction:*

If a proposition $P(n)$ is defined for all positive integral values of n in such a way that $P(1)$ is valid and the validity of $P(k)$ implies that of $P(k+1)$ for an arbitrary positive integer k, then $P(n)$ is valid for all positive integral values of n.

We digress here and consider an example of this principle, even though technically some of the operations and symbols, such as n^2, have not yet been defined. Suppose the proposition $P(n)$ is

$$1 + 3 + 5 + \cdots + (2n - 1) = n^2.$$

For $n = 1$ we have $P(1)$: $1 = 1$, which is valid. Next we let k be any positive integer such that $P(k)$ is valid, i.e.,

$$1 + 3 + 5 + \cdots + (2k - 1) = k^2,$$

and, by adding $2k + 1$ to both sides of this equation, prove the validity of $P(k + 1)$:

$$1 + 3 + \cdots + (2k - 1) + (2k + 1) = k^2 + (2k + 1) = (k + 1)^2.$$

Then, by the principle of mathematical induction, the above proposition $P(n)$ is valid for all positive integral values of n. Other examples of the use of this principle may be found in the following set of exercises.

EXERCISES

Use Peano's postulates to prove the statements in Exercises 1–3; use the principle of mathematical induction to prove the relations in Exercises 4–9.

1. If $a \neq b$, then $a^+ \neq b^+$.
2. $a^+ \neq a$.
3. Every positive integer $a \neq 1$ is of the form b^+, where b is a positive integer.
4. $2 + 4 + 6 + \cdots + 2n = n(n + 1)$.
5. $3 + 6 + 9 + \cdots + 3n = \dfrac{3n(n + 1)}{2}$.
6. $1^2 + 2^2 + 3^2 + \cdots + n^2 = \dfrac{n(n + 1)(2n + 1)}{6}$.
7. $x - y$ is a factor of $x^n - y^n$ where n is any positive integer.
8. $x^{2n} - y^{2n}$ is divisible by $x + y$ where n is any positive integer.
9. $x^{mn} - y^{mn}$ is divisible by $x^m - y^m$ where m and n are any positive integers.

1–5 Addition and multiplication. Peano's postulates are not sufficient to define addition and multiplication explicitly, but they may be used to prove that each of these operations may be defined in exactly one way to satisfy certain conditions. For example, it can be proved [31; 4–5] that given any two positive integers a, b, a positive integer $a + b$ can be defined in exactly one way, so that

$$a^+ = a + 1$$

for every positive integer a, and

$$a + b^+ = (a + b)^+$$

for every pair of positive integers a, b. These two properties of addition and Peano's postulates may then be used to prove that addition of positive integers is unique, associative, and commutative, i.e.,

(i) if a and b are positive integers, there exists a unique positive integer c such that $a + b = c$,
(ii) $(a + b) + c = a + (b + c)$, and
(iii) $a + b = b + a$.

We shall indicate the procedures used to prove these properties. A thorough discussion of this subject may be found in [31; 3–8].

We first prove that $a + b$ is unique for all positive integers a, b. Let a be an arbitrary but fixed positive integer, and S the set of positive integers b such that $a + b$ is a uniquely defined positive integer. From the definition and Peano's second postulate, 1 is in S, since $a^+ = a + 1$ is a uniquely defined positive integer. If b is in S, then $(a + b)^+$ is uniquely defined (Peano's second postulate), and b^+ is in S since $a + b^+ = (a + b)^+$, using the second property in the definition of addition. Thus S contains all positive integers, and $a + b$ is a uniquely defined positive integer for all positive integers a, b.

The proof that addition is associative is similarly obtained (Exercise 1 below) by showing that the set R of positive integers c such that $(a + b) + c = a + (b + c)$ contains all positive integers.

The proof that addition is commutative employs the principle of complete induction twice. We first prove that $a + 1 = 1 + a$ for all positive integers a, and then that $a + b = b + a$ for all positive integers a, b. Let T be the set of positive integers a such that $a + 1 = 1 + a$. The integer 1 is in T since $1 + 1 = 1 + 1$. If a is in T, then $a + 1 = 1 + a$ and, since the proof for the associativity of addition has been indicated, we may apply the second property in the definition of addition twice and obtain $a^+ + 1 = (a + 1) + 1 = a + (1 + 1) = a + 1^+ = (a + 1)^+ = (1 + a)^+ = 1 + a^+$. Thus T contains a^+ and, by the principle of complete induction, the set T contains all positive integers. Finally, let a be an arbitrary fixed positive integer, and U be the set of positive integers b such that $a + b = b + a$. We have just proved that 1 is in U. For any positive integer b in U, we have $a + b^+ = a + (b + 1) = (a + b) + 1 = 1 + (a + b) = 1 + (b + a) = (1 + b) + a = (b + 1) + a = b^+ + a$, whence U contains all positive integers (Exercise 2) and addition has been proved to be commutative.

We have now used Peano's postulates and the two properties of addition in the definition of addition to prove that the addition of two positive integers is unique, associative, and commutative. All five addition properties will be extensively used in the development of our number system. For example, we may now prove the cancellation law for addition, i.e., that $a + b = a + c$ implies $b = c$ (Exercise 3).

Our treatment of multiplication is very similar to that of addition. It can be proved [31; 14–15] that, given any two positive integers a, b, a positive integer written $a \cdot b$, also frequently written ab, can be defined in exactly one way, such that

$$a \cdot 1 = a$$

for every positive integer a, and

$$a \cdot b^+ = ab + a$$

for every pair of positive integers a, b. These two properties of multiplication and Peano's postulates may now be used to prove that multiplication of positive integers is unique (Exercise 6), distributive with respect to addition (Exercises 8 and 11), associative (Exercise 9), and commutative (Exercises 7 and 10). We can also prove the cancellation law for multiplication, i.e., that $ab = ac$ implies $b = c$ for arbitrary positive integers a, b, c (Exercise 12). An integer d is said to have a *factor* or *divisor* b if and only if there exists an integer a such that $d = ab$. The cancellation law for multiplication thus states that if $d = ab$ and $d = ac$, then $b = c$.

We define the *exponential notation* a^b for arbitrary positive integers a, b to indicate the product $aaa \ldots a$ of b factors a. Under this definition $a^b \cdot a^c = a^{b+c}$, $a = d$ implies $a^b = d^b$ (Exercise 16), and $a^b = d^b$ implies $a = d$ (Exercise 17), where a, b, d are arbitrary positive integers.

The definition $a \cdot 1 = a$ and the cancellation law for multiplication imply that if $ab = a$, then $b = 1$. The property $a \cdot 1 = 1 \cdot a = a$ of the positive integer 1 (Exercise 7) is indicated by stating that 1 is *unity*, the *identity for multiplication*. In general, the *identity with respect to an operation* is an element that, when applied to any number of a given set under the given operation, leaves that number unchanged. We shall be particularly concerned with the identity elements for addition and multiplication.

If there existed a positive integer b such that $a + b = a$, then we would have $(a + b)^+ = a^+$, $a + b^+ = a + 1$, and $b^+ = 1$, contrary to Peano's third postulate. Thus there is no identity for addition in the set of positive integers. Accordingly, we now extend the concept of "number" to include a symbol that is not a positive integer by defining a new symbol 0, called *zero*, such that $a + 0 = a$ and $a \cdot 0 = 0$, where a is any positive integer or zero. We shall call zero an integer, but it is not a positive integer. Then (Exercise 18) except for the cancellation law for multiplication, the basic properties of addition and multiplication hold for the set of numbers consisting of the positive integers and zero, i.e., the set of *nonnegative integers*.

The property $a \cdot 0 = 0$ for any nonnegative integer a and the principle of mathematical induction may be used to prove that $0^b = 0$ for any positive integer b (Exercise 15). The symbol 0^0 is *undefined*, i.e., it does not have a specific meaning in our number system. For any positive integer a we define $a^1 = a$ and $a^0 = 1$ in order to retain the property $a^b a^c = a^{b+c}$ for all nonnegative integers b, c.

The equality, addition, and multiplication of positive integers have now been considered. If a and b are positive integers, then $a = c$, $a + b = d$, and $a \cdot b = e$ are also positive integers, i.e., the set of positive integers is closed (Section 1–2) under these operations. In other words, the equations $a = x$, $a + b = y$, and $a \cdot b = z$ all have solutions in the set of positive integers. We have already seen that the solution of $a + x = a$ is not a positive integer. Before defining new numbers to include the solutions of $ax = b$ and $a + x = b$, we shall consider a special case of the second problem. In particular, we shall consider an ordering $a < b$ of the positive integers such that $a < b$ and $b > a$ whenever $a + x = b$ has a solution in the set of positive integers.

Exercises

1. Prove that the addition of positive integers is associative.
2. In the above proof that addition is commutative, give a reason for each step in the proof that U contains all positive integers.
3. Prove that $a + b = a + c$ implies $b = c$ for arbitrary positive integers a, b, c.
4. Prove that if $a = b$, then $a^+ = b^+$ and $a + c = b + c$, where c is any positive integer.
5. Prove that $a = b$ and $c = d$ imply $a + c = b + d$.
6. Prove that ab is a unique positive integer for arbitrary positive integers a, b.
7. Prove that $a \cdot 1 = 1 \cdot a = a$ for all positive integers a.
8. Prove that multiplication is left distributive with respect to addition, i.e., $a(b + c) = ab + ac$.
9. Prove that multiplication is associative, i.e., $(ab)c = a(bc)$.
10. Prove that multiplication is commutative, i.e., $ab = ba$.
11. Prove that multiplication is distributive with respect to addition, i.e., $a(b + c) = ab + ac = (b + c)a$.
12. Prove that $ab = ac$ implies $b = c$ for arbitrary positive integers a, b, c.
13. Prove that $a = b$ implies $ac = bc$, where c is any positive integer.
14. Prove that $a = b$ and $c = d$ imply $ac = bd$.
15. Prove that $0^b = 0$ for any positive integer b.
16. Prove that $a = d$ implies $a^b = d^b$, where a, b, d are arbitrary nonnegative integers and b is any positive integer.
17. Prove that $a^b = d^b$ implies $a = d$, where a, d are arbitrary nonnegative integers and b is any positive integer.
18. Consider the set of nonnegative integers and prove that addition is unique, associative, and commutative, and that multiplication is unique, associative, commutative, and satisfies the distributive law.

1–6 Order relations. Cardinal numbers have been ordered using essentially the definition that $a < b$ if and only if there exists a cardinal number c such that $a + c = b$ (Section 1–2). We now define a similar ordering for the set of positive integers and zero, i.e., for the set of nonnegative integers. Given any two nonnegative integers a, b, we say that *a is less than b* ($a < b$) and *b is greater than a* ($b > a$) if and only if there exists a positive integer c such that $a + c = b$. Then $0 < b$ for every positive integer b (Exercise 1), and $1 < b$ for every positive integer $b \neq 1$ (Exercise 2).

Given any two nonnegative integers a, b, we may now consider three binary relations $a < b$, $a = b$, $a > b$. Let T be the set of nonnegative integers a such that exactly one of the relations $a < b$, $a = b$, $a > b$ holds for every nonnegative integer b. As mentioned in Section 1–3, we shall assume that given any two integers a, b, exactly one of the relations $a = b$, $a \neq b$ holds. For $a = 0$ we have $0 = b$ when $b = 0$ and $0 < b$ when $b \neq 0$. For $a = 1$ we have $b < 1$ if $b = 0$, $1 < b$ if $b \neq 0$ and $b \neq 1$. For any integer a in T we have $b < a^+$ if $b < a$ or $b = a$, $b = a^+$ if $a < b$ and $a + 1 = b$, $a^+ < b$ if $a < b$ and $a^+ \neq b$. Thus, using the principle of mathematical induction, all nonnegative integers are in the set T. In other words, given any two nonnegative integers a, b, exactly one of the relations $a < b$, $a = b$, $a > b$ is valid.

The above definition of $a < b$ for nonnegative integers will be used to define $a < b$ for positive rational numbers (Section 1–8), for negative numbers (Section 1–9), and for real numbers (Section 1–11). The ordering of the real numbers is easily visualized in terms of the one-to-one correspondence between the set of real numbers and the set of points on a line in ordinary Euclidean geometry. This one-to-one correspondence is indicated by the Cantor-Dedekind Axiom (Section 1–12). It also provides a basis for the concept of a linearly ordered set.

A set of elements is *linearly ordered* if for arbitrary elements a, b in the set:

(i) $a \neq b$ implies $a < b$ or $b < a$,
(ii) $a < b$ implies $a \neq b$, and
(iii) $a < b$ and $b < c$ imply $a < c$.

The proof that the set of nonnegative integers is linearly ordered is left as an exercise for the reader (Exercise 3). It can also be proved (Exercise 4) that exactly one of the relations $a < b$, $a = b$, $a > b$ must hold if a and b are elements of any linearly ordered set.

The relation $<$ has several additional properties when a, b, c, d are arbitrary nonnegative integers. For example,

(iv) $a < b$ implies $a + c < b + c$,
(v) $a < b$ and $c < d$ imply $a + c < b + d$,
(vi) $0 < c$ and $a < b$ imply $ac < bc$,
(vii) $a < b$ and $c < d$ imply $ac < bd$,
(viii) $1 < a$ and $b \neq 0$ imply $1 < a^b$,
(ix) $d \neq 0$ and $a < b$ imply $a^d < b^d$,
(x) $a < b$ and $1 < d$ imply $d^a < d^b$, and
(xi) $a < b$ and $1 < c < d$ imply $c^a < d^b$.

The proofs of these properties for nonnegative integers are given as an exercise (Exercise 5). We shall consider the validity and necessary modification of these properties as our concept of number is extended and the binary relations $=$ and $<$ are defined for positive rational numbers, negative numbers, and real numbers.

Exercises

1. Prove $0 < b$ for every positive integer b.
2. Prove $1 < b$ for every positive integer $b \neq 1$.
3. Prove that the set of nonnegative integers is linearly ordered.
4. Prove that exactly one of the relations $a < b$, $a = b$, $a > b$ must hold if a, b are elements of a linearly ordered set.
5. Prove properties (iv) to (xi) of the relation $<$ for nonnegative integers.
6. A set of elements is said to be *well ordered* if every *nonempty* subset (i.e., every subset that contains at least one element) has a first element. Prove that the set of nonnegative integers is well ordered, i.e., prove that if a subset of the nonnegative integers contains at least one element, then it contains an element b such that $b \leqq n$ for every element n of the subset.

1–7 Inverses. We shall use the basic properties of previously considered relations and operations to extend our set of numbers and operations. This will be done by introducing "inverse numbers" and "inverse operations." The inverse of a number n must be considered with reference to a binary operation (Section 1–2) such as addition or multiplication. The numbers 2 and $\frac{1}{2}$ are said to be inverse to each other under multiplication, since $2 \cdot (\frac{1}{2}) = 1$ and 1 is the identity for multiplication (Section 1–5). Also, since $2 + (-2) = 0$, we say that 2 and -2 are inverses under addition. In general, two numbers a, a' are said to be *inverse elements* under an arbitrary operation $\oplus$ with identity element p if and only if $a \oplus a' = p$.

The adjective "inverse" may also be applied to binary operations. Two operations may be called inverse operations if they are opposite in effect, that is, if their successive application with the same number leaves the original number unchanged. For example, $(5 + 2) - 2 = 5$ and also $(5 \cdot 2) \div 2 = 5$. Accordingly, we shall say that subtraction is the inverse of addition and division is the inverse of multiplication. In general, two binary operations $\oplus$ and $\ominus$ are said to be *inverse operations* if and only if $(a \oplus b) \ominus b = a$, where a and b are arbitrary elements of some set of elements over which the operations are defined. We use this relationship and the property that for $b \neq 0$, $ab = cb$ if and only if $a = c$ (Exercises 12 and 13, Section 1–5) to define *division*. We write $a \div b = c$ if and only if $a = bc$. Similarly, for *subtraction* we write $a - b = c$ if and only if $a = b + c$ (see Exercises 3 and 4, Section 1–5).

The relationships among inverse numbers and inverse operations will also be useful in our development of a number system. For example, $5 - 2 = 5 + (-2)$ and $5 \div 2 = 5 \cdot (\frac{1}{2})$. In general, given any element b and inverse elements a, a' under an arbitrary operation $\oplus$ with inverse $\ominus$ such that the operations are both defined for the given elements, we have the relationship $b \oplus a = b \ominus a'$.

The four rational operations, addition, subtraction, multiplication, and division have now been introduced. Addition and multiplication are governed by the properties stated for them in Section 1–5; subtraction and division (excluding division by zero) have been defined as the inverses of addition and multiplication respectively. We may also consider a short form of repeated multiplication, namely, *involution* (the raising of a quantity to a given power), together with its inverse operation *evolution* (the extraction of roots). The need for defining new symbols as numbers, i.e., for gradually expanding the set of elements under consideration, can be seen in terms of these operations. The set of positive integers is closed under addition, multiplication, and involution. Positive rational numbers are needed when division is considered; positive and negative rational numbers and zero when division and subtraction are considered. A still larger set of numbers is needed when evolution is considered. Addition, subtraction, multiplication, division, involution, and evolution can be defined for positive integers in the set of real numbers. We shall consider the set of complex numbers in order to obtain a set of numbers such that these six operations may be defined for all nonzero elements of the set (instead of just for the positive integers).

Let us now return from the above perspective view of our number system to the set of positive integers. The set of positive integers is closed under addition and multiplication. It is not closed under subtraction and division. From a practical viewpoint we may use the positive integers for counting objects or comparing finite sets of objects. We do not yet have numbers to represent such things as, for example, the portion one person receives when three apples are divided equally among six people, or the temperature relative to that at which water freezes. We must, therefore, extend our set of numbers to include fractions (Section 1–8) and signed or directed numbers (Section 1–9). That is, we need inverse numbers for the positive integers under multiplication and addition, along with numbers to represent sums of these new numbers.

Exercises

Prove each of the following for arbitrary nonnegative integers q, r, s, t:

1. $r < s < t$ implies $t - s < t - r$.
2. $r < s < t$ implies $s - r < t - r$.
3. $q < r < s < t$ implies $s - r < t - q$.

1–8 Positive rational numbers. The inverse number under multiplication of the positive integer b is defined to be a number b' satisfying the relation $bb' = 1$. We say that $b' = 1/b$ is the *solution* or *root* of the equation $bx = 1$. It is also called the *zero* of the polynomial $bx - 1$. We now define a new set of numbers, the *positive rational numbers*, so that we may solve equations of the form $bx = a$ for any positive integers a and b. These numbers may be represented by pairs a/b of positive integers with the following properties:

(i) $\dfrac{a}{b} = \dfrac{c}{d}$ if and only if $ad = bc$,

(ii) $\dfrac{a}{b} + \dfrac{c}{d} = \dfrac{ad + bc}{bd}$,

(iii) $\dfrac{a}{b} \cdot \dfrac{c}{d} = \dfrac{ac}{bd}$, and

(iv) $\dfrac{a}{b} < \dfrac{c}{d}$ if and only if $abd^2 < b^2cd$.

The condition $abd^2 < b^2cd$ in (iv) may be stated in the form $ad < bc$ whenever bd is positive. The form $abd^2 < b^2cd$ is used here since it will remain valid when our set of numbers is extended (Section 1–9) to include negative numbers.

Positive rational numbers of the form $a/1$ are identified with the integers a. Technically, the set of positive integers is *isomorphic* to the set of positive rational numbers of the form $a/1$, i.e., there exists a one-to-one correspondence of $a/1$ to a that is preserved under addition and multiplication [$a/1 + b/1$ corresponds to $a + b$ and $(a/1)(b/1)$ corresponds to ab]. This particular isomorphism also preserves order relations (since $a/1 < b/1$ if and only if $a < b$) and is called an *order-isomorphism.*

Equal pairs of integers as specified in (i) above are said to represent the same rational number. In the set of all pairs of integers that are equal to a given pair, there exists one pair, say a/b, such that if r/s is any other pair in the set, then $r = ta$ and $s = tb$ for some positive integer t. A formal proof of the existence of the pair a/b may be given (Exercise 16), using the fact that the set of positive integers is well ordered (Exercise 6, Section 1–6). The pair a/b is said to be the *reduced form* of the given rational number.

The above definitions enable us to prove that for positive rational numbers, addition is unique, associative, and commutative, and that multiplication is unique, associative, commutative, and satisfies the distributive law, i.e., addition and multiplication have the same basic properties (Section 1–5) in the set of positive rational numbers as in the set of positive integers. For example, the sum in (ii) above is unique, since $ad + bc$ and bd are unique for arbitrary positive integers a, b, c, d. Similarly, from

$$\frac{a}{b} + \frac{c}{d} = \frac{ad + bc}{bd} = \frac{cb + da}{db} = \frac{c}{d} + \frac{a}{b},$$

addition is commutative. The remaining proofs are given as exercises at the end of this section.

If e and f are positive rational numbers, then, as in the case of positive integers, $e - f = g$ is defined as a positive rational number if and only if there exists a positive rational number g such that $e = f + g$. Similarly, $e/f = h$ is defined as a positive rational number if and only if there exists a positive rational number h such that $e = fh$. The number g exists if and only if $f < e$ [see Exercise 5(a)]; the number h always exists (Exercise 6). Thus pairs e/f of positive rational numbers are themselves positive rational numbers and nothing new can be obtained by considering in this way pairs of rational numbers instead of pairs of integers. The property [Exercise 5(b)] that for any two positive rational numbers e, f satisfying

$e < f$ there exists a positive rational number r satisfying $e < r < f$ is indicated by saying that the positive rational numbers are *dense*, i.e., between any two distinct positive rational numbers there exists a third positive rational number. The positive integers are not dense.

The symbol a^b has been defined (Section 1–5), where a and b are nonnegative integers and at least one is different from zero. The symbol r^b, where r is any positive rational number and b is any nonnegative integer, may be defined exactly as in the case of integers, i.e., $r^0 = 1$ and r^b for any positive integer b indicates the product of b factors r. We now specify that the symbol $r^{1/b}$ for any positive rational number r and any positive integer b must satisfy the relation $(r^{1/b})^b = r$, i.e., the product of b factors $r^{1/b}$ must equal r. In this manner we shall preserve the property $a^b a^c = a^{b+c}$ for the new symbols. The symbol $r^{d/b}$, where $r > 0$, indicates the product of d factors $r^{1/b}$, and the product $r^s r^t$ is expressed by the symbol r^{s+t} for any positive rational numbers r, s, t. When s is not an integer, the new symbols r^s may not be rational numbers and relations among them are not yet formally defined.

The solutions of the equations $x = a + b$, $x = ab$, and $ax = b$ are positive rational numbers for any positive rational numbers a, b, i.e., the set of positive rational numbers is closed under addition, multiplication, and division. We have extended our concept of number to include $a + b$, ab, and a/b where a, b are arbitrary positive integers or arbitrary positive rational numbers. We next extend our concept of number and define new symbols as numbers, to include $a - b$ where a, b are arbitrary positive rational numbers or zero, i.e., *nonnegative rational numbers.*

Exercises

1. Prove that addition of positive rational numbers is associative.
2. Prove that multiplication of positive rational numbers is (a) unique, (b) associative, (c) commutative, and (d) distributive with respect to addition.
3. Prove that $ac/bc = a/b$, where a, b, c are arbitrary positive integers.
4. Prove that $(a + c)/b = a/b + c/b$, where a, b, c are arbitrary positive integers.
5. Prove that if $a/b < c/d$, where a, b, c, d are arbitrary positive integers, then there exist (a) a positive rational number g such that $a/b + g = c/d$, and (b) a positive rational number r such that $a/b < r < c/d$.

6. Given any positive integers a, b, c, d, prove that there exists a positive rational number h such that $a/b = (c/d) \cdot h$.

7. Define the *arithmetic mean* or *average* of any two positive rational numbers a, b to be $m = (a + b)/2$, and when $a < b$, prove that $a < m < b$.

8. Define $0/1 = 0$, assuming that properties (i) to (iv) above hold for pairs of nonnegative integers a/b, $b \neq 0$, and prove that $0 < r$ for any positive rational number r.

9. Prove that $a < b$ implies $1/b < 1/a$, where a, b are (a) any positive integers, (b) any positive rational numbers.

10. Prove that $a < b$ and $c < d$ imply $a/d < b/c$, where a, b, c, d are (a) any positive integers, (b) any positive rational numbers.

11. Prove that a/b and b/a are inverse numbers under multiplication for arbitrary positive integers a, b.

12. Give the inverse under multiplication whenever such exists for each of the following numbers: 3, 5, 1/10, 1, 0, 10, 200.

13. Using examples, discuss the sets of numbers needed (a) to solve linear and quadratic equations, (b) to perform the six operations discussed in Section 1–7.

14. Prove that the positive rational numbers are linearly ordered (Section 1–6).

15. Restate the eleven properties of the relation $<$ in Section 1–6 to obtain eleven properties for $<$ when a, b, c, d are nonnegative rational numbers.

16. Prove that every positive rational number may be expressed in reduced form.

1–9 Negative numbers. We have just considered pairs a/b of positive integers as symbols, we have defined equality, addition, and multiplication of these symbols, we have proved that these definitions are consistent with the previous definitions over a subset of elements $a/1$ isomorphic with the original set of elements a, and we have obtained new numbers (positive rational numbers) by accepting all symbols a/b as numbers where a and b are positive integers. We now repeat this process by considering pairs $[a - b]$ of nonnegative rational numbers as symbols, by defining equality, addition, and multiplication of these new symbols, by proving the set of symbols $[a - 0]$ isomorphic with the set of nonnegative numbers a under the new definitions, and by then accepting all pairs of nonnegative rational numbers $[a - b]$ as rational numbers — positive, negative, and zero.

We first define the following relations and operations upon the new symbols where a, b, c, d are arbitrary nonnegative rational numbers:

(i) $[a - b] = [c - d]$ if and only if $a + d = b + c$,
(ii) $[a - b] + [c - d] = [(a + c) - (b + d)]$,
(iii) $[a - b] \cdot [c - d] = [(ac + bd) - (bc + ad)]$,
(iv) $[a - b] < [c - d]$ if and only if $a + d < b + c$.

As before, we have defined the meaning of the basic relations $=$, $<$ and the basic operations $+$, $\cdot$ with respect to the new symbols $[a - b]$. We now prove that the correspondence of $[a - 0]$ to a is an order-isomorphism (Section 1–8) and that the basic operations have their usual properties in the complete set of number pairs $[a - b]$.

The correspondence of $[a - 0]$ to a is clearly one-to-one. It remains to prove that the correspondence is preserved under addition, multiplication, and the order relations. These are easily verified since, by definition,

$$[a - 0] + [b - 0] = [(a + b) - (0 + 0)],$$
$$[a - 0] \cdot [b - 0] = [(ab + 0) - (0 + 0)],$$
$$[a - 0] < [b - 0] \text{ if and only if } a + 0 < b + 0.$$

Thus the set of symbols $[a - 0]$, where a is a nonnegative rational number, is equivalent in the sense of being isomorphic to the set of nonnegative rational numbers.

Now let a, b in the symbol $[a - b]$ be arbitrary nonnegative rational numbers. As in the case of the positive rational numbers a/b, we can prove that addition of the symbols $[a - b]$ is unique, associative, and commutative (Exercise 1), and that multiplication is unique, associative, commutative, and satisfies the distributive law (Exercise 2). Accordingly, we consider all symbols of the form $[a - b]$ as numbers, *signed rational numbers*, where a, b are arbitrary nonnegative rational numbers. Under the above isomorphism the signed rational number $[a - b]$ corresponds to 0 when $a = b$ and to d when $a > b$ and $a = b + d$. Since exactly one of the relations $a < b$, $a = b$, $a > b$ must hold (Exercise 4, Section 1–6, and Exercise 14, Section 1–8), we need a similar corresponding symbol for $[a - b] = [0 - c]$ when $a < b$ and $a + c = b$. Accordingly, we now define the symbol $[0 - c]$ to be a *negative rational number* and write $[0 - c]$ as $-c$. The positive rational number c is called the *numerical value* or *absolute value* of $-c$ and of c, i.e., $c = |-c| = |c|$. The positive and negative rational numbers together with zero constitute the signed rational numbers or simply the *rational numbers*. When a, b are positive integers or zero, the pairs $[a - b]$ may be used as

above to define *negative integers*. The positive and negative integers together with zero constitute the *integers*.

Addition and multiplication have been defined for all rational numbers. We now define subtraction and division. Subtraction may be defined for arbitrary rational numbers by the relation

$$[a - b] - [c - d] = [(a + d) - (b + c)]$$

(Exercise 3). Division of rational numbers may be defined by

$$\frac{[a - b]}{[c - d]} = \left[\frac{ac + ad}{c^2 - d^2} - \frac{bc + bd}{c^2 - d^2}\right]$$

when $c \neq d$, is undefined when $c = d$ (Exercises 4 and 5). These definitions enable us to prove that

(i) $-a < -b$ if and only if $b < a$,
(ii) $a + (-b) = a - b$,
(iii) $(-a)b = a(-b) = -ab$, and
(iv) $(-a)(-b) = ab$,

for all nonnegative rational numbers a, b. For example, $-a < -b$ is by definition the same as $[0 - a] < [0 - b]$ and, again by definition, this holds if and only if $0 + b < 0 + a$, that is, $b < a$. Similarly, $a + (-b) = [a - 0] + [0 - b] = [(a + 0) - (0 + b)] = [(a - 0) - (b - 0)] = [(a - b) + (0 - 0)] = a - b$. The remaining two proofs constitute Exercise 6. The proofs indicate that the above properties of signed numbers are consequences of our definitions.

The exponential notation may now be extended to include negative exponents. We define a^b as in Section 1–8 for any positive integer b and any rational number a. Also, as before, $a^0 = 1$ for any $a \neq 0$. A definition for negative exponents is obtained by requiring that a^{-b} satisfy $a^{-b}a^b = 1$ for any integer b and any nonzero rational number a. In this way the property $a^b a^c = a^{b+c}$ is retained for all rational numbers $a \neq 0$ and all integers b, c. The symbol a^b for nonintegral values of b is still formally undefined. It will be considered in Section 1–12. The symbol 0^b is undefined when b is negative or zero.

The set of all rational numbers is closed under addition, subtraction, multiplication, and division (excluding division by zero), i.e., the sum, difference, product, and quotient (divisor different from zero) of any two rational numbers are rational numbers. Many people are primarily concerned with rational numbers. Probably they are sufficient for most grocers, clerks, and even bankers. Yet

the rational numbers also have definite limitations. For example, the distance in feet from home plate to second base in baseball, and the diameter in inches of a baseball nine inches in circumference cannot be exactly stated as rational numbers. They can be expressed approximately as rational numbers, and the error in the approximation can be made less than any given (in advance) positive rational number.

The need for extending the set of rational numbers can also be expressed in terms of magnitudes. We have defined finite sets and finite cardinal numbers (Section 1–2). We now define a number d to be a *finite number* if and only if there exists a positive integer N such that $-N < d < N$. Any magnitude, quantity, object, symbol, etc. is said to be finite if it can be represented by or represents a finite number. We need and shall consider a set of numbers, the set of real numbers, that may be used to compare the magnitudes of any two similar finite objects. Every finite magnitude may be represented as a real number.

Exercises

1. Prove that addition of signed rational numbers is unique, associative, and commutative.
2. Prove that multiplication of signed rational numbers is unique, associative, commutative, and satisfies the distributive law.
3. Prove that subtraction of rational numbers may be defined by $[a - b] - [c - d] = [(a + d) - (b + c)]$ by showing that under this definition it is the inverse of addition.
4. Prove that in the set of positive and negative rational numbers (zero is thus excluded) division as defined above is the inverse of multiplication.
5. Prove that division by zero cannot be defined in the set of rational numbers.
6. Prove the above properties (iii) and (iv) of rational numbers.
7. Prove that $-c < 0$ for any positive rational number c.
8. Prove that $a < b$ and $c < 0$ imply $bc < ac$.
9. Indicate which of the exercises in Section 1–7 may be proved when q, r, s, t are arbitrary rational numbers (positive, negative, or zero).
10. Use the notation $a^{-n} = 1/a^n$ and show that for integers q, r and rational numbers s, t the relations $q < r < 0$ and $0 < s < t < 1$ imply $1 < t^r$, $t^r < t^q$, and $t^r < s^q$.
11. Give the inverse under addition for each of the following numbers: $3, -5, 1/10, 1, 0, 10, -200$.
12. List the rational numbers that are their own inverses under addition, under multiplication.

13. Prove that $[a - b]$ and $[b - a]$ are inverses under addition for arbitrary positive rational numbers a, b.

14. Prove that the rational numbers are linearly ordered.

15. Restate the eleven properties of the relation $<$ in Section 1–6 to obtain eleven properties for $<$ where a, b, c, d are rational numbers.

1–10 Real numbers. The number associated with an object or set of objects usually represents a measure or magnitude relative to some known standard, for example, the height of a tree in feet, the size of a farm in acres, or the number of apples in a box (as compared with one apple). When the measure of one object relative to another cannot be expressed as a quotient of integers, the two objects are said to be *incommensurable.* The early Greeks observed that the diagonal and side of a square are incommensurable. The circumference and diameter of a circle are also incommensurable. Any number that cannot be expressed as the quotient of two integers is said to be *irrational.*

We now prove that $\sqrt{2}$ is irrational. Suppose $\sqrt{2} = a/b$, where a and b are integers with no common integral factors. Then $a^2 = 2b^2$, whence a^2 is an even integer. Therefore, since only an even integer may have an even integer as its square, a is divisible by 2. Let $a = 2c$. Then $4c^2 = 2b^2$, $2c^2 = b^2$, and b is divisible by 2, contrary to our assumption that a and b have no integral factors in common Our initial supposition that $\sqrt{2} = a/b$ is therefore impossible, and $\sqrt{2}$ is irrational.

The above method of proof is often called the method of *indirect proof* or *reductio ad absurdum.* It consists of an assumption that the desired conclusion is false and the use of this assumption and of the given hypothesis in a logical proof of some statement that is contrary to the assumption or to one of the hypotheses (Exercise 1). Then it is said that since the assumption led to a contradiction, the assumption must be false, i.e., the desired conclusion must be true. One form of an indirect proof of a theorem (say, A implies B) is given by a direct proof of the *contrapositive* theorem ("not B" implies "not A"). The method of indirect proof may also be considered as a special case of proof by elimination (Exercise 4).

Irrational numbers such as $\sqrt{2}$ may be defined in several ways. We momentarily digress from our formal development of the number system to discuss briefly the decimal notation for representing all real numbers (rational and irrational). Formally, definitions of and operations with infinite decimals are based upon the very fundamental concepts of infinite sequences and limits (Section 3–11). For the present our considerations will be somewhat intuitive. Formal definitions will be made in the next section in terms of Dedekind cuts. All intuitive concepts used in this section can be rigorously proved from the formal definitions.

We shall prove in Section 2–6 that any positive integer N may be expressed to the "base" 10 in the form

$$N = d_k 10^k + d_{k-1} 10^{k-1} + \cdots + d_2 10^2 + d_1 10 + d_0,$$

where the d_j are elements of the set 0, 1, 2, . . . , 9 of *digits* for the base ten. For example, 1953 means $1 \cdot 10^3 + 9 \cdot 10^2 + 5 \cdot 10 + 3$. Certain fractions may be expressed in the form

$$N + a_1/10 + a_2/10^2 + \cdots + a_m/10^m,$$

where N and m are positive integers and the a_j are digits. For example, $123/4 = 30 + 7/10 + 5/10^2 = 30.75$. Specifically, a rational number $r = a/b$ can be represented using a finite number of terms as above in the decimal notation (i.e., as an *exact decimal*) if and only if r is an integer or $10^m r$ may be expressed as an integer for some positive integer m. Since $2^0 = 5^0 = 1$, this condition can be expressed as follows: a rational number r can be expressed as an exact decimal if and only if there exist integers a, p, q such that $r = a/(2^p 5^q)$.

In order to express the rational number $\frac{4}{3}$ in decimal notation, one must assume that the symbol 1.333 . . . is a number. Formally, this involves the concepts of infinite sequences and limits. Decimals such as the above or $\frac{1}{7}^5 = 2.142857142857 \ldots$, consisting of sets of digits, such as 3 in the case of $\frac{4}{3}$ and 142857 in the case of $\frac{1}{7}^5$, repeated over and over are called *infinite periodic decimals.* Exact decimals may also be considered as infinite periodic decimals by taking $a_j = 0$ for j sufficiently large. For example, $0.25 = 0.2500000 \ldots$. We shall prove in Section 2–7 that every rational number may be represented as an infinite periodic decimal and, conversely, every infinite periodic decimal represents a rational number. The converse statement is easily visualized in terms of the following procedure: Given any periodic decimal d in which a set of k digits is repeated over and over, compute $10^k d - d$ and divide by $10^k - 1$. For example, if $d = 1.333 \ldots$, we compute $10d - d = 13.333 \ldots - 1.333 \ldots = 12$, whence $d = \frac{4}{3}$. If $d = 0.164545 \ldots$, then $100d - d = 16.29$, whence $d = 16.29/99 = 1629/9900$. As mentioned above, a formal definition of the subtraction of infinite decimals requires the concept of limits.

We now define the symbol

$$N + a_1/10 + a_2/10^2 + \cdots + a_n/10^n + \cdots,$$

where N is an integer and the a_i are digits, to be an *infinite decimal.* The above discussion indicates that, after suitable definitions have been made, it is possible to prove that a subset (the infinite periodic decimals) of the set of infinite decimals is isomorphic to the set of rational numbers. In general, one may define equality, sums, and products of infinite decimals so that all infinite decimals behave like numbers. In this way one may represent new numbers, *irrational numbers*, such as $\pi = 3.1415926536\ldots$ [40; 39–40], as *infinite nonperiodic decimals.* The set of all infinite decimals, i.e., the total set of rational and irrational numbers, is called the set of *real numbers.* Accordingly, it is possible to obtain the set of real numbers by assuming that *all infinite decimals represent numbers.* Since this assumption ultimately involves limits of infinite sequences, we shall base our formal development of the real number system on Dedekind cuts (Section 1–11). Section 1–11 is designated as optional to indicate that any reader wishing to assume the usual properties of infinite decimals without formal proof may omit the section.

The real numbers may be classified in several ways. They are positive, negative, or zero. They are rational or irrational. They are algebraic or transcendental. A number is said to be *algebraic* if it satisfies some equation of the form

$$a_0x^n + a_1x^{n-1} + \cdots + a_n = 0, \quad a_0 \neq 0$$

where the a's are integers and n is a positive integer. All other real numbers are said to be *transcendental.* Any rational number a/b satisfies $bx - a = 0$ and is therefore algebraic. Some irrational numbers, such as $\sqrt{2}$ satisfying $x^2 - 2 = 0$, are algebraic. There also exist algebraic numbers, such as i and $-i$ satisfying $x^2 + 1 = 0$, that are not real numbers. A real algebraic number may be rational or irrational; all real transcendental numbers are irrational. The base of natural logarithms

$$e = \lim_{x \to 0} (1 + x)^{1/x}$$

is a real transcendental number; so is π, the ratio of the circumference to a diameter of a circle [29; 71–89, 111]. The symbol πi, where i satisfies $x^2 + 1 = 0$, represents a transcendental number that is not a real number.

Exercises

1. Two statements are *contrary* if they cannot both be true. For example, the statements "The car is a Ford" and "The car is a Dodge" are contrary. Give five pairs of contrary statements.

2. Two statements are *contradictory* if they cannot both be true and also they cannot both be false. The statements given in the illustrative example in Exercise 1 are not contradictory since they might both be false. Contradictory statements are important, since if either one is true, the other is false and if either one is false, the other is true. Give five pairs of contradictory statements.

3. Indicate which of the pairs of statements given in the answer for Exercise 1 are contradictory.

4. The method of proof by elimination consists of considering all possibilities and eliminating all but one of them. For example, if we wish to prove that triangle ABC is a right isosceles triangle, we might consider the possibilities (a) triangle ABC is not a right triangle, (b) triangle ABC is not an isosceles triangle, (c) triangle ABC is a right isosceles triangle. Give another example of this type of reasoning.

5. Prove that $\sqrt{5}$ is an irrational number.

6. Express $1.41414\ldots$ as a rational number.

7. Express $3.176176176\ldots$ as a rational number.

8. Give five rational algebraic numbers.

9. Give five irrational algebraic numbers.

10. Give three numbers that you think are transcendental numbers (formal proof that a given number is transcendental may be very difficult).

11. Indicate the relationship between the classification of real numbers as rational or irrational numbers and the classification of real numbers as exact, infinite periodic, or infinite nonperiodic decimals.

12. Demonstrate our need for extending the real number system by giving five algebraic numbers that are not real numbers.

13. Make a chart indicating the relationships among real, algebraic, transcendental, rational, irrational, and integral numbers.

14. Find equations (with integral coefficients) satisfied by each of the following numbers:

(a) $3 + \sqrt{2}$ (c) $\sqrt{10 - 2\sqrt{5}}$

(b) $\sqrt{3 + \sqrt{2}}$ (d) $\sqrt[5]{4 - \sqrt[3]{2}}$

Are the answers to this exercise unique? Explain.

15. Prove that

$$(a + b\sqrt[3]{c - d})/e$$

is an algebraic number when a, b, c, d, and $e \neq 0$ are integers.

***1–11 Postulates for real numbers.** We now repeat the procedure of defining operations and relations upon new symbols. The new symbols will be called *Dedekind cuts* or simply *cuts*. We shall speak of the following postulate of the existence of numbers corresponding to cuts as the *Dedekind postulate.*

Let the set of all rational numbers be divided into subsets L and R in such a way that every rational number belongs either to L or to R but not to both, neither L nor R is void (empty), and a in L and b in R imply $a < b$. Then there exists a cut number c such that a in L implies $a \leqq c$ and b in R implies $c \leqq b$.

This process of forming a cut may be used on any set of elements in which the order relations (Section 1–6) are defined. We shall be primarily concerned with cuts $\{L, R\}$ in the set of rational numbers. If L consists of all rational numbers $x \leqq 3$, R of all $x > 3$, then $c = 3$ and c is in L. If L consists of all $x < 5$, R of $x \geqq 5$, then $c = 5$ and c is in R. Note that c is rational and is in L or R in both of these examples. If L consists of all negative x and of all nonnegative x such that $x^2 < 2$, and if R consists of all positive x such that $x^2 > 2$, then all rational numbers are in L or R and $c = \sqrt{2}$ is in neither L nor R (Section 1–10). In general, when the sets L and R are subsets of the set of rational numbers, the cut number c is in L or R if, and only if, c is rational. A cut is said to be *closed* if the cut number c is an element of the set, *open* otherwise. Thus a cut in the set of rational numbers is closed if c is rational, open if c is irrational, i.e., not rational. The set of all cut numbers obtained from cuts in the set of rational numbers is called the set of *real numbers.* The effect of performing a cut in the set of real numbers is usually stated as a theorem, the *Dedekind Theorem: Every cut in the real number system is closed* (Exercise 10).

Given any two cuts $\{L, R\}$ and $\{S, T\}$ in the set of rational numbers, we now make the following definitions:

(i) $\{L, R\} = \{S, T\}$ if there is at most one element of S that is not an element of L and, conversely, there is at most one element of L that is not an element of S,

(ii) $\{L, R\} < \{S, T\}$ if there are at least two elements of S that are not elements of L,

* The asterisk indicates that this section may be omitted without disturbing the organization of the text.

(iii) $\{L, R\} + \{S, T\} = \{U, V\}$ where U consists of all rational numbers that can be expressed in the form $a + s$, a in L, s in S.

The "at most one element" in (i) and "at least two elements" in (ii) are necessary since the cut $\{L, R\}$ where L consists of all $x < 2$ and the cut $\{S, T\}$ where S consists of all $x \leqq 2$ have the same cut number $c = 2$ and must be considered equal. Each of the above definitions can also be expressed (Exercise 4) in terms of conditions upon R, T, and V, since by definition of a cut $\{I, R\}$ in the set of rational numbers every rational number must be in L or R, whence R consists of all rational numbers that are not in L.

The cut $\{N, P\}$, where all negative rational numbers and zero are in N and all positive rational numbers are in P, is called a *zero cut.* A cut $\{L, R\}$ is said to be *negative* if $\{L, R\} < \{N, P\}$, *zero* if $\{L, R\} = \{N, P\}$, *positive* if $\{N, P\} < \{L, R\}$. A cut that is positive or zero is said to be *nonnegative.* The product of two cuts $\{L, R\} \cdot \{S, T\} = \{U, V\}$ may be defined by considering the possible cases of negative and nonnegative cuts. We define V to be the set of rational numbers expressible in the form as, a in L, s in S, when the two given cuts are both negative; the set of rational numbers expressible in the form rt, r in R, t in T, when both given cuts are nonnegative. When one of the given cuts is nonnegative and the other is negative, U consists of the set of rational numbers expressible in the form at, a in L, t in T, when $\{L, R\}$ is negative; the set of rational numbers expressible in the form sr, s in S, r in R, when $\{S, T\}$ is negative.

We have now defined equality, order relations, addition, and multiplication of the new symbols $\{L, R\}$. As in the case of symbols a/b and $[a - b]$, it remains to prove that there exists an order-isomorphism between a subset of the new symbols and the given set of numbers, the set of rational numbers. We shall consider the correspondence of $\{L, R\}$ to c where c is a rational number and L consists of all rational numbers $x \leqq c$. This is essentially the correspondence between closed cuts and rational numbers, since if V consists of all rational numbers $\geqq c$, then $\{U, V\} = \{I, R\}$.

Given two cuts $\{L, R\}$, $\{S, T\}$ corresponding to rational numbers a and b respectively, the above definitions may be used to prove that $\{L, R\} = \{S, T\}$ if and only if $a = b$, $\{L, R\} < \{S, T\}$ if and only if $a < b$, $\{I, R\} + \{S, T\}$ corresponds to $a + b$, and $\{L, R\} \cdot \{S, T\}$ corresponds to $a \cdot b$. The proofs are not difficult and are left as an

exercise (Exercise 7) for the reader. The order-isomorphism between the set of closed cuts and the set of rational numbers shows that for the set of closed cuts, the above definitions are consistent with our previous definitions. As mentioned above, the cut numbers indicated by the new symbols are called real numbers for arbitrary cuts $\{L, R\}$ in the set of rational numbers. In this manner we have obtained new numbers, *irrational numbers*, corresponding to open cuts. The Dedekind Postulate for the existence of a cut number c for all cuts in the set of rational numbers serves as a postulate for the existence of all real numbers, rational and irrational, in terms of rational numbers. The Dedekind Theorem (all cuts in the set of real numbers are closed) may be proved (Exercise 10) by showing that every cut in the set of real numbers defines a cut in the set of rational numbers. Several other properties of Dedekind cuts and real numbers will be considered in the following exercises and in Section 1–12.

Exercises

1. Give two examples of open cuts in the set of rational numbers.
2. Give two examples of closed cuts in the set of rational numbers.
3. Show that every cut in the set of integers is closed.
4. Rephrase definitions (i), (ii), and (iii) above in terms of R, T, and V.
5. Give a zero cut $\{N', P'\} = \{N, P\}$ such that the sets N' and N are different.
6. Construct the cut that is the sum of the two cuts given in the answer for Exercise 1.
7. Given two closed cuts $\{L, R\}$, $\{S, T\}$ corresponding to a and b respectively, prove that $\{L, R\} = \{S, T\}$ if and only if $a = b$, $\{L, R\} < \{S, T\}$ if and only if $a < b$, $\{L, R\} + \{S, T\}$ corresponds to $a + b$, and $\{L, R\} \cdot \{S, T\}$ corresponds to $a \cdot b$.
8. Repeat Exercise 6 for the product.
9. Define subtraction of cuts and repeat Exercise 6 for subtraction.
10. Prove the Dedekind Theorem.
11. Define division of an arbitrary cut by a cut different from zero. Give a numerical example.
12. Prove that the real numbers are linearly ordered (Section 1–6).
13. Prove that the real numbers are dense (Section 1–8).

1–12 Properties of real numbers. We now assume that the real numbers exist and are linearly ordered (Exercise 12, Section 1–11). They may be considered either as decimals (Section 1–10) or as

Dedekind cuts (Section 1–11). We also assume that the four rational operations have been defined and have the same properties in the set of real numbers as in the set of rational numbers (Section 1–11).

Given any real number a and any positive integer b, we define the symbol a^b to represent the product of b factors a. When $a \neq 0$, we define $a^0 = 1$ and $a^{-b}a^b = 1$ for any integer b. When $a = 0$ and b is any positive real number, $a^b = 0$. We define $(a^{1/b})^b = a$ for any rational number b when $a > 0$ and for any odd integer b when $a < 0$. These definitions preserve the property $a^b a^c = a^{b+c}$ whenever the symbols are defined. In general, for any real number a and any rational number b, say $b = r/s$ where the integers r, s have no common factors, the symbol a^b represents a real number if and only if the integer r may be chosen so that a^r is defined and the equation $x^s = a^r$ has a solution in the set of real numbers. This concept may be used (Exercises 10, 11, 12, and 13) for real numbers a to give precise meaning to the symbol a^b in the set of real numbers under any of the following conditions:

(i) $a = 0$ and b is any positive real number,
(ii) $a > 0$ and b is any real number, and
(iii) $a < 0$ and $b = r/s$ where the integers r, s have no common factors and s is odd (does not have 2 as a factor).

When $a = 0$ and $b \leqq 0$, we cannot give explicit meaning in the set of real numbers to the symbol a^b and at the same time preserve the property $a^b a^c = a^{b+c}$. When $a < 0$ and b is irrational, or $b = r/s$ where r and s have no common factors and s is even, the symbol a^b does not have an explicit meaning in the set of real numbers but can be defined in the set of complex numbers (Section 1–16).

All properties of the set of real numbers may also be considered as properties of the set of points on a line. We shall assume that the reader is familiar with the use of a rectangular Cartesian coordinate system in Euclidean plane geometry. In any such given coordinate system we shall define points having integers as coordinates to be *integral points*, points having rational numbers as coordinates to be *rational points*, and points having real numbers as coordinates to be *real points*. Given an origin and a unit point, all integral and rational points may be constructed with straightedge and compasses (Section 6–4). Some irrational points such as $\sqrt{2}$, the diagonal of a unit square, may also be constructed. The existence of the set of

irrational points must be postulated. Euclid assumed that any line segment joining the center of a circle to a point outside the circle contained a point of the circle. We shall assume the *Cantor-Dedekind Axiom:*

> *To each point of the line there corresponds one and only one real number and, conversely, to each real number there corresponds one and only one point of the line.*

This one-to-one correspondence may be chosen, as in the case of the usual coordinate systems, so that there is an order-isomorphism between the set of points on the line and the set of real numbers. The correspondence then makes possible a geometric visualization of the properties of the set of real numbers. For example, the real and rational numbers are dense (Section 1–8); the integers are not dense. The integers, rational numbers, and real numbers are each linearly ordered (Section 1–6).

The property that distinguishes the set of rational numbers from the set of real numbers is continuity. This is the property used in plane geometry to prove that any line joining the center of a circle to a point outside the circle must cut the circle in at least one point. Intuitively, a line or curve is continuous if there are no "holes" in it. Technically, we may use the Cantor-Dedekind Axiom, represent the elements of any given linearly ordered set as an ordered set of points on a line, and consider the real numbers (coordinates) associated with these points. We then define the given linearly ordered set of elements and the associated set of points to be *continuous* if and only if the corresponding set of real numbers includes all real numbers x or includes all real x satisfying one of the relations $a < x$, $x < b$, $a < x < b$ for some real numbers a, b. This definition can be stated much more elegantly by using the terminology of Section 1–11. A linearly ordered set of elements is said to be continuous if it is dense and satisfies the Dedekind Postulate. Thus the rational numbers are dense but not continuous; the real numbers are dense and continuous. The above definitions of dense and continuous sets may be extended to sets of points in a plane and to many other sets that cannot be linearly ordered.

The statements listed in Exercise 9 essentially designate methods of extending a linearly ordered dense set to a continuous set. For example, each of the numbers

$$1, 1.4, 1.41, 1.414, 1.4142, 1.41421, 1.414214, \ldots$$

may be expressed as a rational number with a power of ten as its denominator (Section 1–10). However, if we consider such a sequence of numbers $x_1, x_2, x_3, \ldots$, satisfying

$$2 - x_1^2 > 2 - x_2^2 > 2 - x_3^2 > \ldots,$$

and such that for any given positive rational number e the positive number $2 - x_n^2$ can be made less than e by choosing n sufficiently large, i.e., making closer and closer approximations to $\sqrt{2}$, then this nonending sequence of numbers may be said to define a new number $\sqrt{2}$. In the language of analysis, the above sequence of rational numbers has $\sqrt{2}$ as a limit (Section 3–11). In general, any linearly ordered dense set of elements may be made continuous by adding all the limits of convergent sequences (Section 3–11) of its elements. For example, the set of rational numbers was made into a continuous set (the set of real numbers) by adding the irrational numbers. Every irrational number may be expressed as a limit of a sequence of rational numbers.

The final property of real numbers that we shall consider is that of boundedness. The word "bound" is frequently used to indicate a limit that cannot or should not be exceeded. The phrase "out of bounds" is common in many games. Every physical object has bounds. The vast polar wastes of Antarctica are bounded by oceans. The air we breathe is a part of the earth's atmosphere and is bounded, since it does not extend to the sun, another planet, or even the moon. The number of hairs on a person's head and the number of grains of sand on Miami Beach are bounded even though, at least in the second case, the number is large.

The word "unbounded" is used to indicate that an object or the elements of a set exceed any bound that may be tried for it. The set of positive integers is said to be unbounded since if any real number M is tried as a bound, there exists an integer n such that $n > M$. To be more exact, we say that the positive integers are bounded below by zero, or any negative number, and are unbounded above. The same may be said for the positive real numbers. The negative integers are unbounded below and bounded above. The set of all integers is unbounded below and unbounded above, i.e., *unbounded.* Similarly, the real numbers are unbounded.

The set of positive real numbers $\leqq N$ is bounded by 0 and N. Any enumerated set of real numbers is bounded. For example, the set 1, 5, 75, 32, 17, -4 is bounded by -4 and 75 or by -10 and

100, ... Thus bounds are not unique, and any specified real number is bounded. For example, any real number n is bounded by $n - 1$ and $n + 1$. Each individual real number is bounded, but the set of all real numbers is unbounded. The same may be said for the set of integers. A set of real numbers is said to be *bounded* if there exists a fixed positive integer N such that $-N < b < N$ for every element b of the set. In the next section we shall consider unbounded sets of elements and be particularly concerned with sets of elements that can be placed in one-to-one correspondence with the set of positive integers.

EXERCISES

1. Indicate which of the following sets of elements are linearly ordered: (a) integers, (b) rational numbers, (c) irrational numbers, (d) real algebraic numbers, (e) points of a circle in Euclidean geometry, and (f) points of a bounded line segment in Euclidean geometry.

2. Indicate which of the sets of elements in Exercise 1 are dense.

3. Indicate which of the sets of elements in Exercise 1 are continuous.

4. Show that every linearly ordered continuous set must also be dense.

5. Give a property of the set of real numbers that distinguishes it from the set of rational numbers.

6. Give a property of the set of rational numbers that distinguishes it from the set of integers.

7. Give three sets of numbers having each of the following properties: (a) bounded below and above, (b) bounded below only, (c) bounded above only, (d) unbounded.

8. Give two sets of bounds, if any exist, for each set of numbers given in the answers for Exercise 7.

9. Give algebraic or geometric examples for each of the following statements. Each statement may be used to postulate the set of real numbers and, in this sense, is equivalent to the Dedekind Postulate. Any one of these statements may also serve as a basis for continuity in both algebra and geometry.

(a) All decimals exist as real numbers.

(b) *Bolzano-Weierstrass Theorem:* Every infinite bounded set has at least one limiting point.

(c) Every set of points that is bounded below has a greatest lower bound.

(d) Every set of points that is bounded above has a least upper bound.

(e) *Heine-Borel-Lebesgue Theorem:* If an infinite set of intervals I covers a fundamental set of points S on a finite closed interval, then there exists a finite subset of I that covers S.

(f) Every Cauchy sequence of rational numbers determines a real number (Section 3–11).

(g) *Cantor Theorem:* Given any sequence of intervals $E_1, E_2, E_3, \ldots$ on a line where E_i is given by $a_i \leqq x \leqq b_i$ and the relations $a_1 \leqq a_2 \leqq a_3 \leqq \ldots$, $\ldots \leqq b_3 \leqq b_2 \leqq b_1$ hold, then there exists at least one point, say $x = x_0$, that is contained in every interval of the sequence.

10. Define a^b for any real number $a \neq 0$ and any integer b.

* 11. Define a^b as a positive number for any positive real number a and any real number b.

12. Define a^b for any negative real number a and any rational number $b = r/s$, where the integers r, s have no common factors and s is odd.

13. Prove that a^b is uniquely defined under the conditions stated in Exercises 10, 11, and 12.

1–13 Transfinite cardinal numbers. Cardinal numbers associated with finite sets of elements have been considered in Sections 1–1 and 1–2. Two finite sets have the same cardinal number if and only if there exists a one-to-one correspondence between the elements of the two sets. We now use one-to-one correspondences to associate cardinal numbers (*transfinite cardinal numbers*) with infinite sets (Section 1–2).

Two infinite sets of elements have the same transfinite cardinal number if and only if there exists a one-to-one correspondence between the elements of the two sets. As in the case of finite sets, two infinite sets having the same cardinal number are said to be equivalent. However, an infinite set may be equivalent to one of its proper subsets. For example, the set of positive integers n is equivalent to the set of even positive integers under the correspondence of n with $2n$. Because of this property of infinite sets, a one-to-one correspondence between the elements of a set A and a proper subset of a set B implies only that the cardinal number of the set A is less than or equal to the cardinal number of the set B. In order to prove that the cardinal number of a set A is less than the cardinal number of a set B, it is necessary to show that A is equivalent to a proper subset of B and that there does not exist a one-to-one correspondence between the elements of A and the elements of B.

If a set of elements has cardinal number three, its elements may be placed in one-to-one correspondence with the set 1, 2, 3 of positive integers. If a set has cardinal number n, its elements may be placed in one-to-one correspondence with the set 1, 2, 3, . . . , n of positive integers. If the elements of a set may be placed in one-to-

* The solution of this exercise involves the material presented in Section 1–11.

one correspondence with the set of all positive integers 1, 2, 3, . . . , its cardinal number is called *aleph-zero*, $\aleph_0$. The set is said to be *countably infinite* or *denumerably infinite*. The set of all positive integers

$$1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad \ldots \quad n \quad \ldots,$$

the set of even positive integers

$$2 \quad 4 \quad 6 \quad 8 \quad 10 \quad 12 \quad 14 \quad \ldots \quad 2n \quad \ldots,$$

the set of odd positive integers

$$1 \quad 3 \quad 5 \quad 7 \quad 9 \quad 11 \quad 13 \quad \ldots \quad 2n-1 \quad \ldots,$$

and even the set of positive rational numbers, as we shall prove, are all countably infinite.

The set of positive rational numbers is at least countably infinite since it has the set of positive integers as a subset. We shall prove that the set of positive rational numbers is at most countably infinite by proving that the set of all pairs of positive integers is countably infinite. Consider the array

$$\begin{array}{cccccccc}
1/1 & 1/2 & 1/3 & 1/4 & 1/5 & 1/6 & 1/7 & \ldots \\
2/1 & 2/2 & 2/3 & 2/4 & 2/5 & 2/6 & 2/7 & \ldots \\
3/1 & 3/2 & 3/3 & 3/4 & 3/5 & 3/6 & 3/7 & \ldots \\
4/1 & 4/2 & 4/3 & 4/4 & 4/5 & 4/6 & 4/7 & \ldots \\
5/1 & 5/2 & 5/3 & 5/4 & 5/5 & 5/6 & 5/7 & \ldots \\
6/1 & 6/2 & 6/3 & 6/4 & 6/5 & 6/6 & 6/7 & \ldots \\
\ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \ldots
\end{array}$$

and associate a positive integer with each pair by going along the diagonal lines of the array as indicated:

$$1 \sim 1/1, \quad 2 \sim 1/2, \quad 3 \sim 2/1, \quad 4 \sim 3/1, \quad 5 \sim 2/2, \quad \ldots.$$

This one-to-one correspondence between the set of positive integers and the set of all pairs of positive integers indicates that the set of pairs is countably infinite. Since the set of positive rational numbers is a subset of the set of all pairs of positive integers, the set of positive rational numbers is at most countably infinite. Then, since it is also at least countably infinite, the set of positive rational numbers is countably infinite.

The above correspondence between the positive rational numbers and the positive integers gives an ordering of the positive rational numbers. This ordering satisfies the conditions for a linearly ordered set (Section 1–6) but is clearly not an ordering according to the magnitude or measure of the numbers. For example, under this ordering 2 precedes both 1/4 and 3.

Since the set of multiples of a thousand, the set of even integers, the set of odd integers, the set of integers, and the set of rational numbers all have the same cardinal number, the question arises as to whether all infinite sets of real numbers are countably infinite. It can be proved that the set of algebraic real numbers is countably infinite (Exercise 5) and that the set of transcendental real numbers is not countably infinite (Exercise 6).

In terms of the points on a line, it is easy to visualize "holes" in the line when only the integral points are present. When all the rational points have been added, the points are dense, and yet, considering the line as the x-axis, the circle of radius $\sqrt{2}$ and center at the origin crosses the line without hitting a rational point, so there must still be "holes" in the line. Even if all the points having algebraic real numbers (Section 1–10) as coordinates were added, there would still be "holes," since, if the circumference of a unit circle could be cut, straightened, and one end put at the origin, the other end would be at a transcendental point, 2π. There are no "holes" in the line when all real points are present.

Consider the set of real numbers between zero and one and let them be represented as decimals. If the set is countably infinite, the decimals can be placed in one-to-one correspondence with the positive integers, and they can be listed in the order imposed by this correspondence. Suppose the imposed order is

$$\begin{array}{l}
.1284\ldots \\
.2315\ldots \\
.1694\ldots \\
.7850\ldots \\
\ldots\ldots \\
\ldots\ldots
\end{array}$$

Under the assumption that the set is countably infinite, all decimals between zero and one are in the above array. Suppose we construct a decimal by taking the elements .1390 . . . on the main diagonal and increasing each element by 1, except that 9 is replaced by 0. The new decimal in this case is .2401 . . . and lies between zero and one. It is not on the first row above since the first elements differ;

it is not on the second row since the second elements differ; and in general it is not on the jth row since the jth elements differ. Thus the constructed decimal is not in the array, and the assumption that the real numbers between zero and one are countable has led to a contradiction. Briefly, if the set of real numbers between zero and one is countably infinite, the real numbers of that set can be listed in some order as above. Whatever this order may be, we may, if necessary, reorder the numbers so that at least one of the digits on the main diagonal is not 8 or 9 and, by the above procedure, construct a new decimal that is not in the list. Thus the real numbers between zero and one cannot be listed in order and the set is not countably infinite. The set of real numbers is then infinite but not countably infinite, since one of its subsets is infinite and not countably infinite.

The cardinal number associated with the set of all real numbers is called the *cardinal number of the continuum*, C. The countable or denumerable infinity, $\aleph_0$, is the first or smallest transfinite cardinal number. One of the famous unsolved mathematical problems is to prove that C is the next transfinite number after $\aleph_0$, i.e., $C = \aleph_1$. There is at least a countably infinite set [13; 84–85], [29; 54–55] of different transfinite cardinal numbers, but the above two are the most common.

The correspondence between real numbers and the points on a line in ordinary Euclidean geometry holds only for finite numbers. Transfinite numbers are not included in the set of real numbers and have relations quite different from those of real numbers. For example, $\aleph_0 \pm a = \aleph_0$, where a is equal to $\aleph_0$ or any finite cardinal number, $\aleph_0 + C = C$, $\aleph_0 \cdot 5 = \aleph_0$.

EXERCISES

1. Give three examples of each of the following: (a) a finite set and a finite proper subset, (b) an infinite set and a finite subset, (c) an infinite set and an infinite proper subset, (d) a countably infinite set, (e) an infinite set that is not countably infinite.

2. Show that the negative rational numbers are countably infinite.

3. Show that any function $f(x)$ defined (Section 3–10) for positive integral values of x takes on a countably infinite sequence of values (not necessarily distinct) as x takes on the values $1, 2, 3, \ldots$.

4. Give three examples of sequences of numbers obtained as indicated in Exercise 3.

5. Prove that the set of algebraic real numbers is countably infinite [13; 103].

6. Prove that the set of transcendental real numbers is not countably infinite.

1–14 Group; number system. We have discussed the rational number system and the real number system. In this section we shall consider the concept of a group and state exactly what is meant by a "number system."

A set of elements forms a *group* with respect to an arbitrary unique binary operation $\oplus$ (Section 1–2) if it is (1) closed, (2) associative, and contains (3) an identity element and (4) the inverse of each of its elements. In other words, the set C of elements $a, b, \ldots$ forms a group with respect to $\oplus$ if (1) $a \oplus b$ is in C for all pairs a, b in C, (2) $(a \oplus b) \oplus c = a \oplus (b \oplus c)$ for all a, b, c in C, (3) there is an element I in C such that $I \oplus a = a \oplus I = a$ for all a in C, and (4) for every a in C there is an a' in C such that $a \oplus a' = a' \oplus a = I$. For example, the set of integers (positive, negative, and zero) forms a group with respect to the operation of addition but not with respect to the operation of multiplication. The set of rational numbers (and also the set of real numbers) forms a group with respect to addition. When zero is excluded, the remaining rational numbers form a group with respect to multiplication.

A group is said to be *commutative* (or Abelian) if $a \oplus b = b \oplus a$ for all elements a, b in the group.

A set of elements in which two binary operations $+$ and $\times$ are defined forms a *number system* or *skew field* if (1) the set forms a commutative group with respect to $+$, (2) the set with the identity for $+$ removed forms a group under $\times$, and (3) the distributive laws of $\times$ with respect to $+$,

$$a \times (b + c) = a \times b + a \times c, \quad (b + c) \times a = b \times a + c \times a,$$

hold for any three elements a, b, c of the set. A number system in which $\times$ is commutative is called a *field.*

If a and b are elements of a number system such that $ab = 0$ and $a \neq 0$, then a^{-1} exists from (2), $a^{-1}ab = 0$, and $b = 0$. In other words, if the product of two elements of a number system is zero, then at least one of the elements must be zero. Formally, an element $k \neq 0$ is called a *zero divisor* if there exists an element $j \neq 0$ such that $j \cdot k = 0$ (see Exercise 13, Section 2–9). The above proof shows that zero divisors cannot exist in a number system.

If a number system contains the positive integers, it must also contain (i) zero and the negative integers, i.e., the identity and inverses for addition, and (ii) the rational numbers, since it must also include the inverses under multiplication of all integers different from zero and all finite sums of these inverses. Thus any number system that contains the positive integers must contain the rational numbers. The rational numbers, real numbers, and, as we shall soon see, complex numbers each form number systems. Since we have defined both addition and multiplication to be commutative, the sets of rational, real, and complex numbers each also form fields.

The above exact definition of a number system provides the basic concepts of this common mathematical term. Of even more fundamental importance is the introduction of the concepts of group and field that, along with ring (Section 1–18), form the basis for most of the definitions in abstract algebra.

EXERCISES

1. Indicate which of the following sets of numbers form groups under addition:
   (a) even integers,
   (b) odd integers,
   (c) integral multiples of ten,
   (d) integral multiples of any integer k,
   (e) positive integers,
   (f) numbers of the form $b\sqrt{2}$, where b is a rational number,
   (g) numbers of the form $a + b\sqrt{2}$, where a and b are integers,
   (h) numbers of the form $a + b\sqrt{2}$, where a and b are rational numbers,
   (i) 0,
   (j) positive rational numbers,
   (k) irrational numbers,
   (l) numbers of the form $a + bw$, where a and b are any rational numbers and w is a given algebraic number (Section 1–18).

2. In each of the sets of numbers in Exercise 1, exclude zero whenever it is present and indicate which of the resulting sets of numbers form groups under multiplication.

3. Prove that if a set of elements is associative under addition and closed under subtraction, it forms a group under addition.

1–15 Complex numbers. We have started with the positive integers, developed the rational number system to obtain a set of numbers that is closed under the four rational operations (addition, multiplication, subtraction, division), and developed the real number system to obtain a set of numbers in which every finite magnitude can be represented. In the rational and real number systems, the relations of order and equality as well as the operations of addition and multiplication have the same basic properties as in the set of positive integers (Sections 1–5 and 1–6). In this section we repeat once more the procedure for defining relations and operations for a new symbol, showing that these definitions are consistent with previous definitions over a subset of the symbols, and defining the new symbols to be numbers. All previously considered finite symbols for numbers could be represented as points on a line in ordinary plane geometry and were linearly ordered. The new symbols that we consider now must be represented on a plane instead of a line in ordinary plane geometry. Thus the new symbols are not linearly ordered and we shall not consider their order relations.

The basic need for the new symbols that we are about to introduce arises from the desirability of solving algebraic equations. Our previous developments of finite algebraic numbers can be considered from the point of view of finding numbers to correspond to all real roots of a polynomial equation, i.e., numbers to designate the points at which a polynomial curve crosses the x-axis in the real plane. When a, b, and c are arbitrary positive integers, the positive rational numbers are needed in order to solve all equations of the form $ax = b$, negative numbers are often needed to solve equations of the form $x + a = b$, and real numbers (rational and some irrational) to solve equations of the form $ax^2 + bx + c = 0$ where $b^2 - 4ac \geqq 0$. All zeros of a polynomial $f(x)$ *that appear as ordinary geometric intersections* of the graph of $y = f(x)$ with the real x-axis are, of course, real numbers. However, algebraically it is desirable that the quadratic equation $x^2 + 2ax + b = 0$ have two roots, whether the curve $y = x^2 + 2ax + b$ intersects the x-axis or not in Euclidean plane geometry, i.e., for any real values of a and b. Accordingly, we again extend our number system to include a new type of number.

We now consider ordered pairs of real numbers (a, b) where

(i) $(a, b) = (c, d)$ if and only if $a = c$ and $b = d$,
(ii) $(a, b) + (c, d) = (a + c, b + d)$,
(iii) $(a, b) \cdot (c, d) = (ac - bd, ad + bc)$.

As in the case of other new symbols, we consider a correspondence between a subset $(a, 0)$ of the new symbols and the set of real numbers

a. Also, as before, the above definitions may be used to prove that this correspondence is an isomorphism (Exercise 1). Note that in this case the isomorphism is not called an order-isomorphism.

The new symbols (a, b) are defined to be numbers, *complex numbers*, for arbitrary real numbers a, b. The number a is called the *real part* of the complex number (a, b); b is called the *imaginary part*. The complex number (a, b) is said to be *imaginary* if $b \neq 0$, *pure imaginary* if $b \neq 0$ and $a = 0$. Under the above isomorphism the set of complex numbers (a, b) consists of the set of real numbers $(b = 0)$ and the set of imaginary numbers $(b \neq 0)$. Since the complex numbers are not linearly ordered according to magnitude, the classification of numbers as positive, zero, and negative is used only for real numbers. Thus negative number and positive number always refer to negative real number and positive real number.

We can now prove that the roots of $x^2 + 1 = 0$ are $(0, 1)$ and $(0, -1)$. We have, in fact, $(0, 1)^2 = (0, 1) \cdot (0, 1) = (0 - 1, 0 + 0) = (-1, 0) = -1$ and $(0, -1)^2 = (0, -1) \cdot (0, -1) = (0 - 1, 0 + 0) = (-1, 0) = -1$. The mechanical manipulation of complex numbers is greatly simplified by writing $(0, 1) = i$. Then $(a, b) = (a, 0) + (0, b) = (a, 0) + (b, 0) \cdot (0, 1) = a + bi$, and we may treat a, b, i as numbers, subject to the condition that $i^2 = -1$ whenever i^2 occurs. Thus,

$$\begin{aligned}(2, 3)^3 = (2 + 3i)^3 &= 8 + 36i + 54i^2 + 27i^3 \\ &= 8 + 36i - 54 - 27i \\ &= -46 + 9i = (-46, 9).\end{aligned}$$

The word *complex* denotes that the new numbers are not simple numbers as we have understood them in the past, but that each is an ordered pair of such numbers satisfying (i) and (ii) above. It is unfortunate that the word imaginary is used as opposed to real. Except in the technical sense agreed upon by mathematicians, the two kinds of numbers are equally real.

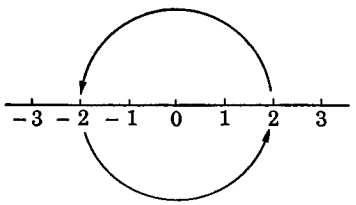


FIG. 1-2

The negative numbers may be considered as arising from a rotation of the positive x-axis about the origin through 180°. If this rotation is applied twice, one obtains the identity $-(-a) = a$. In this sense, multiplication by -1 and rotation through 180° are equivalent

(Fig. 1-2). Similarly, multiplication by i is equivalent to a 90° rotation. If the multiplication or the rotation is applied twice, the negative number is obtained and if applied four times, the original number is obtained (Fig. 1-3). A plane such as that in Fig. 1-3 with an axis of real numbers and an axis of imaginary numbers is often called a *complex plane*. Each complex number $a + bi = (a, b)$ may be associated with a unique point in the complex plane, with the coordinate a along the real axis and b along the imaginary axis (Section 1-16).

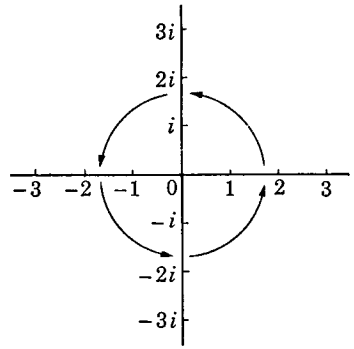


FIG. 1-3

The complex numbers $a + bi$ and $a - bi$ are each the *conjugate* of the other. The *norm* $n(z)$ of a complex number z is the product of the number and its conjugate. Thus, if $z = a + bi$, $n(z) = n(a + bi) = (a + bi)(a - bi) = a^2 + b^2$, which, for $z \neq 0$, is always positive.

The *absolute value*, or *modulus*, of $z = a + bi$ is the nonnegative square root of the norm, $|z| = \sqrt{a^2 + b^2}$. Thus the absolute value of a complex number is always a real number.

When we seek the quotient of two complex numbers $(a, b) \div (c, d)$, we actually seek a number (p, q) such that

$$\cdot(a, b) = (c, d) \cdot (p, q).$$

We have

$$\begin{aligned}(a, b) &= (cp - dq, cq + dp), \\ a &= cp - dq, \\ b &= dp + cq,\end{aligned}$$

whence

$$p = \frac{ac + bd}{c^2 + d^2}, \quad q = \frac{bc - ad}{c^2 + d^2}$$

if $c^2 + d^2 \neq 0$. Thus from the uniqueness of the sum, difference, product, and quotient of real numbers we have unique real numbers p and q whenever $c^2 + d^2 \neq 0$, i.e., whenever $(c, d) \neq 0$. This completes the proof of the following theorem.

THEOREM 1–1. *In the complex number system, division, except by zero, is always possible and is unique.*

In practice, the division of $z_1 = a + bi$ by $z_2 = c + di$ is customarily indicated by the quotient $z_1/z_2 = (a + bi)/(c + di)$. This quotient is then expressed as a complex number by multiplying its numerator and denominator by the conjugate of z_2, that is,

$$\frac{a+bi}{c+di} = \frac{(a+bi)(c-di)}{(c+di)(c-di)} = \frac{ac+bd}{c^2+d^2} + \frac{bc-ad}{c^2+d^2}\,i.$$

Other representations of z_1/z_2 are considered in Section 1–16. At that time we shall also be concerned with relations between the absolute values or moduli of z_1 and z_2. In particular, we shall need the fact that the absolute value of a product of complex numbers is equal to the product of the absolute values of its factors. This fact is a consequence of Theorem 1–2, which we now state and prove.

THEOREM 1–2. *The norm of a product is equal to the product of the norms of its factors.*

Let $z_1 = a + bi$, $z_2 = c + di$, then $n(z_1) = a^2 + b^2$, $n(z_2) = c^2 + d^2$, $z_1z_2 = ac - bd + (ad + bc)i$, and

$$\begin{aligned} n(z_1z_2) &= (ac - bd)^2 + (ad + bc)^2 \\ &= a^2c^2 + b^2d^2 - 2abcd + a^2d^2 + b^2c^2 + 2abcd \\ &= (a^2 + b^2)(c^2 + d^2) \\ &= n(z_1) \cdot n(z_2). \end{aligned}$$

This proves the theorem for the product of two factors. Since the product of two complex numbers is itself a complex number, the proof may be reapplied with $z_1 = w_1w_2$, $z_2 = w_3$ to prove the theorem for three factors, and in general with $z_1 = w_1w_2 \ldots w_{n-1}$ and $z_2 = w_n$ to prove the theorem for any finite number of factors by mathematical induction (Section 1–4).

In the case of addition of complex numbers we may prove

THEOREM 1–3. *The absolute value of a sum of complex numbers is less than, or equal to, the sum of the absolute values.*

Suppose $|z_1 + z_2| > |z_1| + |z_2|$ where $z_1 = a + bi$, $z_2 = c + di$. Then

$$\sqrt{(a+c)^2 + (b+d)^2} > \sqrt{a^2 + b^2} + \sqrt{c^2 + d^2},$$

whence

$$\begin{aligned} (a+c)^2 + (b+d)^2 &> a^2 + b^2 + 2\sqrt{(a^2+b^2)(c^2+d^2)} + c^2 + d^2, \\ a^2 + c^2 + b^2 + d^2 + 2ac + 2bd &> a^2 + b^2 + c^2 + d^2 + 2\sqrt{(a^2+b^2)(c^2+d^2)}, \\ ac + bd &> \sqrt{(a^2+b^2)(c^2+d^2)}, \\ a^2c^2 + b^2d^2 + 2abcd &> a^2c^2 + b^2d^2 + a^2d^2 + b^2c^2, \\ 0 &> b^2c^2 - 2abcd + a^2d^2, \\ 0 &> (bc - ad)^2. \end{aligned}$$

But this is impossible, since the right side is the square of a real number, and therefore nonnegative. Thus the assumption that the theorem is false has led to a contradiction, and we have given an indirect proof (Section 1–10) of the theorem for the sum of two complex numbers. This proof can be extended by mathematical induction to cover any finite sum, just as the preceding proof for products was extended.

EXERCISES

1. Prove that the correspondence of $(a, 0)$ to a is an isomorphism.
2. Use your previous knowledge and give a quadratic equation with real coefficients that has its roots in the set of (a) integers, (b) rational numbers, (c) real irrational numbers, (d) imaginary numbers, (e) pure imaginary numbers.
3. Prove that the complex numbers form a number system.
4. Express in the form $a + bi$:

$$\sqrt{-16}, \quad 5, \quad -\tfrac{2}{3}, \quad 1 + \sqrt{-2}, \quad \frac{2+\sqrt{-3}}{2-\sqrt{-3}}, \quad \frac{1}{3+4i}.$$

5. Determine the modulus of each of the complex numbers given in Exercise 4.
6. Prove that if a complex number is equal to its conjugate, the number is real.
7. Prove that if the product of two complex numbers is zero, then at least one of the numbers is zero.

1–16 Properties of complex numbers. The relationships between algebra and geometry are especially important to anyone endeavoring to learn the fundamental concepts of mathematics. We have seen (Section 1–12) that all real numbers may be represented as points on a line and, conversely, all points on a line in Euclidean geometry may be represented by real numbers. In this section we consider two representations of complex numbers on a plane in Euclidean

geometry. Then from these graphic representations we derive trigonometric and exponential representations for complex numbers.

Given a rectangular Cartesian coordinate system with origin O and axes Ox and Oy, we take Ox as the axis of reals and Oy as the axis of imaginaries. The complex number $z = a + bi$ may be represented either by the point P:(a, b) or, if $z \neq 0$, by the directed line segment, *vector*, **OP** (Fig. 1–4). A vector has both length and direction. The length r of **OP** is given by the absolute value of z; its direction is given by the angle θ between the positive x-axis and **OP**. For each z, the nonnegative number $r = \sqrt{a^2 + b^2}$ is uniquely determined, but $\theta = \arctan b/a$, the *amplitude* or *argument* of z, can be determined only to within a multiple of 2π. The complex number $z = a + bi$ and the point P: (a, b) are uniquely determined by either the pair of real numbers (a, b), that is, the *Cartesian* or *rectangular coordinates* of the point P, or by the pair of real numbers (r, θ), that is, the *polar coordinates* of the point P. In terms of trigonometric functions, $a = r\cos\theta$, $b = r\sin\theta$, and $z = r(\cos\theta + i\sin\theta)$. Any complex number z may also be expressed by using exponential notation.

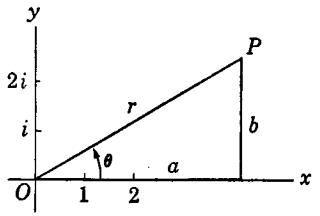


FIG. 1–4

For real numbers a, b the symbol a^b may be defined as a unique real number (Exercise 13, Section 1–12) when $a = 0$ and $b > 0$, when $a > 0$ and b is any real number, and when $a < 0$ and $b = r/s$, where the integers r, s have no common factors and s is odd. The unique value of the symbol a^b is based upon the fact that for rational values of $b = r/s$, where r, s are integers without common factors, the equation $x^s = a^r$ has a unique positive solution in the set of real numbers when $a^r > 0$, a unique real solution in all other cases such that a^b is defined. Since the equation $x^s = a^r$ has s solutions in the complex number system, the symbol $a^{r/s}$ may be associated with any one of s complex numbers. Thus, in our definition of a^b in the complex number system, it will at times be necessary to designate a particular element from a subset of the complex numbers as the *principal value* of a given symbol a^b.

The exponential representation $z = re^{i\theta}$, where e is the base for natural logarithms, may be derived from the trigonometric representation $z = r(\cos\theta + i\sin\theta)$, using infinite series. The readers who do not recall the following infinite series from their previous work in

mathematics may review the development of these series in Section 3–15 or take the representation $z = re^{i\theta}$ as an additional assumption.

The infinite series

$$e^x = 1 + x + \frac{x^2}{2} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

may be used to define e^x for any complex number x (Section 3–15). We then obtain

$$e^{ix} = 1 + ix - \frac{x^2}{2} - \frac{ix^3}{3!} + \frac{x^4}{4!} + \frac{ix^5}{5!} - \cdots$$

upon substituting ix for x. A comparison of this series with the two series

$$\cos x = 1 - \frac{x^2}{2} + \frac{x^4}{4!} - \frac{x^6}{6!} + \cdots,$$

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots$$

indicates that $e^{ix} = \cos x + i\sin x$. Thus we have three representations for a complex number, $z = a + bi = r(\cos\theta + i\sin\theta) = re^{i\theta}$. The conditions for the equality of two complex numbers z_1, z_2 are $a_1 = a_2$, $b_1 = b_2$ when the numbers are expressed in the first form. The conditions are $r_1 = r_2$, $\theta_1 = \theta_2 + 2k\pi$ for some integer k when the other two forms are used. In general, we shall find the first form most useful when considering sums of complex numbers, one of the other forms when considering products or powers.

The sum of $z_1 = a + bi$ and $z_2 = c + di$ has been defined (Section 1–15) to be $z_1 + z_2 = a + c + (b + d)i$. To find geometrically the sum of two complex numbers z_1, z_2 represented by P_1 and P_2, respectively, construct the parallelogram having OP_1 and OP_2 as sides (Fig. 1–5). The diagonal OS of the parallelogram is the vector representing $z_1 + z_2$. The difference $z_2 - z_1$ may be constructed as a side OD of a parallelogram with diagonal OP_2 and one side OP_1 (Fig. 1–6).

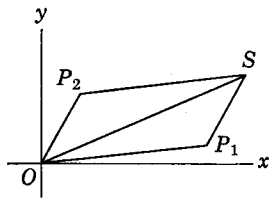


FIG. 1–5

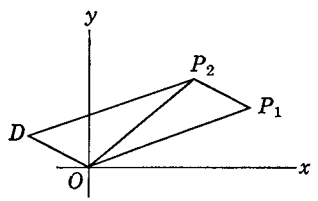


FIG. 1–6

The product z_1z_2 is defined to be $z_1z_2 = ac - bd + (ad + bc)i$, but is much more readily interpreted in the form $z_1z_2 = r_1e^{i\theta_1} \cdot r_2e^{i\theta_2} = r_1r_2e^{i(\theta_1+\theta_2)}$. The form $z_1z_2 = r_1r_2[\cos(\theta_1 + \theta_2) + i \sin(\theta_1 + \theta_2)]$ can be verified trigonometrically. Then, using mathematical induction as in Theorem 1–2, we have (Exercise 8)

THEOREM 1–4. *The absolute value of the product of two or more complex numbers is the product of their absolute values; the amplitude of the product is the sum of their amplitudes.*

To find geometrically the product of z_1, z_2 represented by P_1, P_2, construct triangle OP_2P similar to triangle OAP_1 where $O = (0, 0)$,

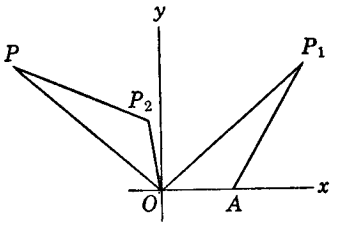


FIG. 1–7

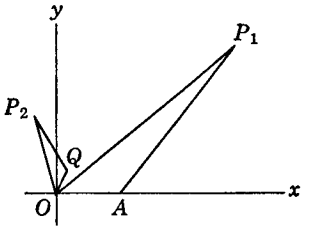


FIG. 1–8

and $A = (1, 0)$ (Fig. 1–7). Then P is the point representing $z_1 \cdot z_2$. In this construction the two triangles OAP_1 and OP_2P must be similarly oriented. For example, if the interior of triangle OAP_1 is on the left as one traverses the perimeter of the triangle from O to A to P_1 to O, then the interior of triangle OP_2P must be on the left as one traverses its perimeter in the sense OP_2P. Triangles OAP_1 and OP_2P are similarly oriented in Fig. 1–7; triangles OAP_1 and OP_2Q are oppositely oriented in Fig. 1–8.

The corresponding geometrical construction for a quotient z_2/z_1 is obtained by constructing similarly oriented triangles OP_2Q and OP_1A (Fig. 1–8). Using $z_2/z_1 = (r_2/r_1)e^{i(\theta_2-\theta_1)}$, we have, corresponding to Theorem 1–4: the absolute value of the quotient of two complex numbers is the quotient of their absolute values; the amplitude of the quotient is the amplitude of the dividend minus the amplitude of the divisor.

As in the case of real numbers, the inverse operations of subtraction and division can be avoided by finding the inverse numbers $-z$ and $1/z$ respectively. If $z = a + bi$, then

$$-z = -a - bi = r[\cos(\pi + \theta) + i \sin(\pi + \theta)] = re^{i(\pi+\theta)},$$

using Theorem 1–4, and $-1 = \cos\pi + i \sin\pi$. Also when $r \neq 0$,

$$1/z = 1/r[\cos(-\theta) + i \sin(-\theta)] = 1/r(\cos\theta - i \sin\theta),$$

using the corresponding statement for quotients, and

$$1 = 1(\cos 0 + i \sin 0).$$

We have now considered the sum, difference, product, and quotient of two complex numbers. With the exception of the order relations, all our previous rules apply in the complex number system. There is no satisfactory definition of the magnitude or measure of a complex number that can be used to order linearly the set of all complex numbers. In fact, we have seen that the complex numbers correspond to the points of a plane instead of a line. Thus the order relations previously considered for real numbers should not be expected to apply for complex numbers. The complex numbers are dense and continuous when the definitions of those terms are stated for nonlinear sets.

The fundamental importance of the complex numbers is based upon the fact that they enable us to state two roots for any quadratic equation $x^2 + 2ax + b = 0$ with real coefficients without restrictions upon the sign of $a^2 - b$. Actually, the n roots of any polynomial equation of degree n (Section 3–1 and Theorem 4–2) with complex coefficients may be expressed as complex numbers (Section 1–18). This property is described by saying that the complex number system is *algebraically closed* [7; 393]. The roots of the equations $w^2 = z$, $w = z^2$ and, in general, $w^n = z$, $w = z^n$ for any positive integer n and any given complex number z have been of particular interest to mathematicians and will be considered in Section 1–17.

EXERCISES

1. Express each of the following complex numbers in the form $r(\cos\theta + i \sin\theta)$:

(a) $1 + i$, (c) 15, (e) $-8 + 8i\sqrt{3}$,
(b) $2 - 2i$, (d) $7i$, (f) -3.

2. Express each of the complex numbers in Exercise 1 in the form $re^{i\theta}$.
3. Add the following pairs of numbers graphically:

(a) $3 - i$, $5 + 2i$, (c) $\sqrt{5} - \sqrt{-1}$, $2 - \sqrt{-16}$,
(b) $3 + \sqrt{-27}$, i, (d) $2 + 2\sqrt{-3}$, $\dfrac{1 + i\sqrt{2}}{2}$.

4. Multiply the pairs of numbers in Exercise 3 graphically and check the answers algebraically.

5. Subtract graphically the first number from the second in each part of Exercise 3 and check the answers algebraically.

6. Divide graphically the first number by the second in each part of Exercise 3 and check the answers algebraically.

7. State the conditions under which each of the following holds:

(a) $|z_1 + z_2| = |z_1| + |z_2|$,
(b) $|z_1 + z_2| = |z_1| - |z_2|$.

8. Prove Theorem 1–4.

1–17 De Moivre's Theorem. Given any complex number $z = re^{i\theta}$ and any positive integer n, we have defined z^n to represent the product of n factors z. Then, by Theorem 1–4, we have $z^n = r^n e^{in\theta}$. Similarly, for any given complex number $z = re^{i\theta}$ and any positive integer n, the n complex roots of the equation $w^n = z$ may be expressed in the form

$$z^{1/n} = (re^{i\theta})^{1/n} = r^{1/n}e^{(\theta+2k\pi)i/n}$$

for $k = 0, 1, 2, \ldots, n-1$, using the fact that θ is determined only to within a multiple of 2π. The values $k = n, n+1, \ldots$ are not used, since the sine and cosine of $(\theta + 2k\pi)/n$ have equal values for $k = n$ and $k = 0$, for $k = n+1$ and $k = 1, \ldots$. However, in both cases, z^n and $z^{1/n}$, the results are most easily obtained and remembered by using the ordinary rules for exponents.

The symbol z^n has a unique value for any positive integer n. The symbol $z^{1/n}$ may take on any one of n values in the complex number system. When z is a real number, these possible values of the symbol $z^{1/n}$ include the real value represented by $z^{1/n}$ in the set of real numbers. This value is called the principal value (Section 1–16) of $z^{1/n}$ in the set of complex numbers. It may be obtained by taking $k = 0$ when z is positive, and $k = (n-1)/2$ when z is negative and n is odd. When z is negative and n is even, we define the principal value of $z^{1/n}$ to be that obtained when $k = 0$. We shall not attempt to designate principal values of $z^{1/n}$ when z is imaginary.

Consider, for example, $z = 2 + 2i\sqrt{3}$ with $r = 4$ and $\theta = 60° = \pi/3$, that is, $z = 4e^{\pi i/3}$. We then have $z^2 = 16e^{2\pi i/3}$ and $z^{\frac{1}{2}} = 2e^{(\pi/3+2k\pi)i/2}$, where $k = 0, 1$. We next use the relation $e^{ix} = \cos x + i \sin x$ (Section 1–16) to express z^2 and $z^{\frac{1}{2}}$ in the form $a + bi$. In particular, $z^2 = 16(\cos 2\pi/3 + i \sin 2\pi/3) = 16(\cos 120° + i \sin 120°) = -8 + 8i\sqrt{3}$.

For $k = 0$, $z^{\frac{1}{2}} = 2(\cos \pi/6 + i \sin \pi/6) = \sqrt{3} + i$; for $k = 1$, $z^{\frac{1}{2}} = 2(\cos 7\pi/6 + i \sin 7\pi/6) = -\sqrt{3} - i$. If we take $k = 2$, then $z^{\frac{1}{2}} = 2(\cos 13\pi/6 + i \sin 13\pi/6) = 2(\cos \pi/6 + i \sin \pi/6)$, and we have the same value of $z^{\frac{1}{2}}$ as for $k = 0$. The following theorem states a general form of these results.

THEOREM 1–5. DE MOIVRE'S THEOREM. *If n is any positive integer and $z = r(\cos\theta + i \sin\theta)$, then*

$$z^n = [r(\cos\theta + i\sin\theta)]^n = r^n(\cos n\theta + i \sin n\theta) = r^n e^{in\theta};$$

$$z^{1/n} = r^{1/n}\{\cos[(\theta + 2k\pi)/n] + i \sin[(\theta + 2k\pi)/n]\}$$

$$= r^{1/n}e^{(\theta+2k\pi)i/n}, \quad k = 0, 1, 2, \ldots, n-1.$$

We use this theorem for any complex number $z = re^{i\theta}$ and positive integer n to express the unique complex number z^n and the n complex roots of the equation $w^n = z$. Each of these n roots has absolute value $r^{1/n}$, that is, each is represented by a point on a circle of radius $r^{1/n}$ with center at the origin. These points are equally spaced on the circle since, when taken in the order of the corresponding values of k, their amplitudes differ by consecutive multiples of $2\pi/n$. In the above example of $z = 2 + 2i\sqrt{3}$, the two roots of $w^2 = z$ had amplitudes of $\pi/6$ and $\pi/6 + 2\pi/2 = 7\pi/6$, respectively, and both had absolute value 2. In general, we have

THEOREM 1–6. *Any complex number $z = r(\cos\theta + i \sin\theta)$ not equal to zero has exactly n distinct complex nth roots which may be represented by n points equally spaced on a circle of radius $r^{1/n}$.*

In particular, for $z = 1$, the cube roots of unity satisfy

$$w^3 = 1 = 1(\cos 0 + i \sin 0)$$

and therefore may be expressed as

$$w = 1^{1/3}[\cos(0 + 2k\pi)/3 + i \sin(0 + 2k\pi)/3]$$

for $k = 0, 1, 2$ or as

$$w_1 = 1(\cos 0 + i \sin 0) = 1,$$
$$w_2 = 1(\cos 120° + i \sin 120°) = -\tfrac{1}{2} + i\sqrt{3}/2,$$
$$w_3 = 1(\cos 240° + i \sin 240°) = -\tfrac{1}{2} - i\sqrt{3}/2.$$

The points representing w_1, w_2, w_3 are the vertices of an equilateral triangle inscribed in a unit circle about the origin and having one vertex at $(1, 0)$ on the positive x-axis. In general, the nth roots of

unity are represented by the vertices of a regular polygon of n sides inscribed in the unit circle, one vertex lying at (1, 0) on the positive x-axis.

Considered from a slightly different point of view, the nth roots of unity form a group (Section 1–14) of n elements. It is called a *cyclic group*, since every element of the group can be expressed in terms of a single element. In the above example, $w^3 = 1$, the three roots could be expressed as w_2, w_2^2, w_2^3, or as w_3, w_3^2, w_3^3. An nth root of unity s is a *primitive* nth *root* of unity if n is the smallest positive integer m such that $s^m = 1$, that is, in the language of group theory the primitive nth roots are those of *order* n.

The nth roots of unity as obtained from $z^n = \cos 0 + i \sin 0 = 1$ are $\cos 2k\pi/n + i \sin 2k\pi/n$ for $k = 0, 1, 2, \ldots, n-1$, by De Moivre's Theorem. In particular, for $k = 1$ the root $w = \cos 2\pi/n + i \sin 2\pi/n$ is a primitive nth root, since $w^t = \cos 2t\pi/n + i \sin 2t\pi/n$ can equal unity if and only if t is a multiple of n, that is, n is the least positive power of w that is equal to unity. Thus there exists at least one primitive nth root of unity for any positive integer n. We now proceed to find all nth roots (not necessarily primitive) from a single primitive nth root of unity.

Given any primitive nth root of unity s and any integer t, we have $(s^t)^n = (s^n)^t = 1^t = 1$, whence any positive integral power of a primitive nth root is also an nth root of unity. Also, if $s^t = s^u$, we may suppose $u \leqq t$ and write $s^{t-u} = 1$. But n is the least positive power of s that equals unity, since s is a primitive nth root. Thus either $t = u$ or $t - u$ is a multiple of n (Section 2–2 and Theorem 2–8), that is, $s^t = s^u$ if and only if $t = u + kn$ for some integer k. Accordingly, we have proved that the numbers $s, s^2, s^3, \ldots, s^n = 1$ are distinct nth roots of unity. Under the assumption (Theorem 4–2) that the polynomial equation $z^n - 1 = 0$ has n roots, we have

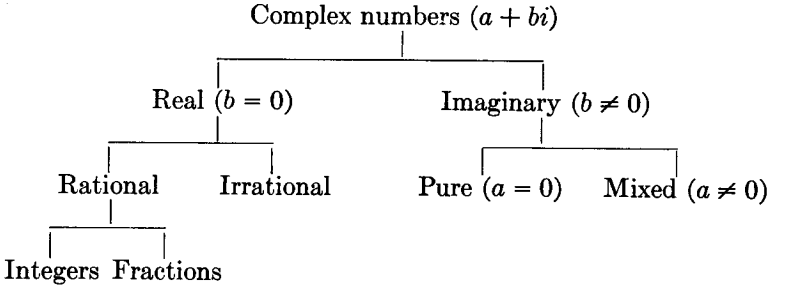
THEOREM 1–7. *If s is a primitive nth root of unity, all nth roots of unity are given by the sequence*

$$s, s^2, s^3, \ldots, s^{n-1}, s^n = 1.$$

We shall later (Theorem 2–17) find that the primitive nth roots of unity are precisely the numbers s^t, where t and n are relatively prime and s is any primitive nth root of unity.

The study of groups is of considerable importance in mathematical theories. We shall consider them further as we discuss a few more

abstract concepts in Section 1–18. So far, we have considered the properties of the positive integers (Section 1–4) and from these developed the rational, real, and complex number systems. The needs for and the advantages of each system have been discussed. With certain identifications such as $a + 0 \cdot i$ with a and $a/1$ with a, we may consider the complex number system to be "our number system" and to have the other sets of numbers as subsets. The following chart indicates some of these subdivisions of the complex number system and the associated conditions upon the real numbers a, b in the expression $a + bi$ for a general complex number.



The real numbers may also be classified as positive, zero, or negative; the complex numbers, as algebraic or transcendental.

The complex number system may thus be considered as the basis for our study of the fundamental concepts of algebra. Our next chapter, the elementary theory of numbers, concerns some special properties of the integers. In particular, we shall consider sufficient properties to prove that every repeating decimal represents a rational number and conversely (Section 2–7). Polynomials have many properties similar to those of the integers and a parallel study of the theory of polynomials is made in Chapter 3 before considering the theory of polynomial equations in Chapter 4.

EXERCISES

1. Without using De Moivre's Theorem, find the square roots of $11 - 60i$, $5 + 12i$, $-i$, $24 + 70i$, $-4ab + (2b^2 - 2a^2)i$.

(*Hint:* Assume $\sqrt{z} = x + iy$, where x and y are to be determined.)

2. Find z by De Moivre's Theorem when $z^4 = 16$, $z^3 = -27$, $z^5 = i$, $z^3 = -8i$, $z^3 = 4 + 4\sqrt{-3}$.

3. Expand $(\cos\theta + i \sin\theta)^3$ by De Moivre's Theorem. Also, expand it by the binomial theorem and thus obtain formulas for $\cos 3\theta$ and $\sin 3\theta$.

4. Prove that the numbers $1, -1, i, -i$ form a group (Section 1–14) under multiplication.

5. Find all the fifth roots of unity and represent them graphically.

6. Find all the values of $\sqrt[5]{1+i}$ and $\sqrt[3]{i}$ and represent them graphically.

7. Expand by De Moivre's Theorem:

$$(2\sqrt{3} - 2i)^5, \quad [4(\cos 150° + i \sin 150°]^4.$$

8. Find the three cube roots of $-27, -i, 1 + i$.

9. If w, w' are the conjugate complex cube roots of unity, show that $1 + w + w^2 = 0$, $w' = w^2$, $w = w'^2$, $w \cdot w' = 1$.

10. Indicate the roots of the equation $z^n - 1 = 0$.

11. Make a chart indicating which of the fourteen sets of numbers (complex, imaginary, pure imaginary, mixed imaginary, real, rational, irrational, integral, fractional, positive, zero, negative, algebraic, transcendental) mentioned above are closed (Section 1–2) under (a) addition, (b) subtraction, (c) multiplication, (d) division.

12. Make a chart as in Exercise 11 indicating which of the fourteen sets of numbers are (a) groups under addition, (b) groups under multiplication after zero has been excluded from each set that contains zero, (c) fields (i.e., commutative number systems).

13. Prove that the nth roots of unity form a cyclic group for any positive integer n.

***1–18 Fields and number systems.** The basis for the similarity mentioned above between the properties of polynomials and integers is found in the fact that both sets form rings (to be defined shortly). In this section we shall consider a few fundamental but somewhat abstract concepts of our number system. The concepts of group, field, and number system have been defined in Section 1–14. It was also shown in Section 1–14 that the set of rational numbers forms the smallest field (commutative number system) that contains the positive integers. In this section we shall consider a development of the complex numbers by adjoining (to be defined) numbers to the sets of rational and real numbers. We shall also observe that the set of complex numbers forms a field in which every polynomial equation in one unknown with coefficients from the complex number system can be factored into linear factors.

We start with the positive integers or natural numbers. In order to form a group under addition we must include the negative integers and zero. The set of all integers forms a ring. In general, a set of elements for which addition and multiplication are uniquely defined forms a *ring* if the elements of the set

(i) form a commutative group under addition,
(ii) are closed under multiplication,
(iii) satisfy the associative law of multiplication, and
(iv) satisfy the distributive law of multiplication with respect to addition (Section 1–14).

The set of even integers satisfies the above definition and forms a ring. Thus there exist rings of numbers which do not contain unity, the identity under multiplication.

A *field* may be defined as a ring in which

(i) there is an identity element, unity, for multiplication,
(ii) multiplication is commutative, and
(iii) every element except zero has an inverse under multiplication.

Thus any set of elements (for example, the set of rational, real, or complex numbers) that forms a field also forms a ring, but not conversely. The integers form a ring but do not form a field. In general, the elements of a ring or field are not necessarily numbers. For example, the set of all polynomials in the indeterminate x with integral coefficients forms a ring; the set of all rational functions in x with integral coefficients forms a field. A comprehensive treatment of groups, rings, and fields may be found in both [7] and [52], and a very readable treatment in [34].

We extended the set of positive integers in Section 1–8 by considering quotients a/b, where a and b were positive integers. In general, we may associate with any ring the quotients a/b where a and $b \neq 0$ are elements of the ring and b is not a zero divisor (Section 1–14). This set of quotients forms a field and is called the *quotient field* of the ring. Thus the quotient field of the ring of integers is the field of rational numbers, R.

The field R may be extended by *adjoining* (as described in the following sentence) an element k which does not belong to R. The extended field is composed of all quotients (denominator different from zero) of polynomials in k with coefficients from R. If k is an element of R, nothing new is obtained. The set of all polynomials in k with coefficients from R forms a ring denoted by $R[k]$. The set of all rational functions of k (quotients of polynomials where the polynomial in the denominator is different from zero) with coefficients from R forms a field denoted by $R(k)$. The field $R(k)$ is called an *algebraic extension* of the field R if k is a root of a polynomial $f(x)$, not identically zero, with coefficients from R. For example, the

ring $R[\sqrt{2}]$ consists of all numbers of the form $a + b\sqrt{2}$, where a and b are from R; the field $R(\sqrt{2})$ also consists of all numbers $a + b\sqrt{2}$, since

$$\frac{c + d\sqrt{2}}{e + h\sqrt{2}} = \frac{c + d\sqrt{2}}{e + h\sqrt{2}} \cdot \frac{e - h\sqrt{2}}{e - h\sqrt{2}} = a + b\sqrt{2}$$

for a suitable choice of a and b in R. In general, the ring $T[k]$, obtained by adjoining a number k that is algebraic over T to a field T, is also a field, that is, $T[k] = T(k)$ when k is algebraic over T. Using terminology which has not been defined in this chapter but which is familiar to nearly everyone, we may prove the above statement as follows: Suppose $f(k) = 0$ and consider $g(k)/h(k)$ where $f(x)$, $g(x)$, and $h(x)$ are polynomials, $f(x)$ is irreducible over T, and $h(k) \neq 0$. Then $f(x)$ and $h(x)$ have highest common factor unity, whence by the Euclidean Algorithm (Section 3–7) there exist polynomials $p(x)$ and $q(x)$ such that

$$p(x)f(x) + q(x)h(x) = 1.$$

Since $f(k) = 0$, we have $q(k)h(k) = 1$, whence

$$\frac{g(k)}{h(k)} = g(k) \cdot q(k)$$

is a polynomial in k, that is, $T[k] = T(k)$.

Finally, given the ring of integers with quotient field R, we seek a field T such that every polynomial $f(x)$ with coefficients from R factors completely into linear factors (each with coefficients from T). Since $f(x)$ may be linear, every element of R is an element of T. Thus T is an extension field of R. Suppose we try to construct T by adjoining numbers to R. In order to factor $x^2 - 2$ one must adjoin $\sqrt{2}$ and obtain $R(\sqrt{2})$. Similarly, for any prime number 2, 3, 5, 7, 11, 13, 17, . . . one must adjoin $\sqrt{p}$ in order to factor $x^2 - p$ into linear factors. In Section 2–3 it will be shown that there are infinitely many prime numbers. Thus no finite number of adjunctions of the form $\sqrt{p}$ to R will suffice even to factor all quadratic expressions of the form $x^2 - p$, and some other approach to the problem is necessary.

One approach to this problem involves the concept of continuity (Section 1–12), and therefore the real number field R^* is needed. The field R^* contains $\sqrt{p}$ for all primes p and thus is sufficient to factor polynomials of the form $x^2 - p$. Since the real numbers are uncountable, we could not expect to obtain R^* by a finite number of algebraic extensions of R.

The field of complex numbers $R^*(i)$ constitutes an algebraic extension of R^*, since i satisfies the equation $x^2 + 1 = 0$. Also, it can be shown that the field $R^*(i)$ cannot be further extended algebraically, i.e., that every polynomial of positive degree with coefficients from $R^*(i)$ has all its roots in $R^*(i)$. This result is usually proved in two steps. First one proves the Fundamental Theorem of Algebra (Theorem 4–3): Every polynomial $p(x)$ of positive degree with complex coefficients has at least one complex zero. Then Theorem 4–2 is proved: Every polynomial of degree $m > 0$ with complex coefficients has precisely m complex zeros (not necessarily distinct). There are several proofs of the Fundamental Theorem of Algebra, and each involves nonalgebraic concepts [7; 114]. Accordingly, the reader is referred to [7; 113–115] and other similar texts for an intuitive proof. In Chapter 4 we shall adopt Theorem 4–3 without proof and use it to prove Theorem 4–2. It can also be proved [7; 393] that every polynomial with algebraic coefficients has all its roots in the set of algebraic numbers (Section 1–10). In other words, any polynomial with algebraic coefficients may be factored into linear factors with algebraic coefficients. Further discussion of the topics considered in this section may be found in most abstract algebra texts.

Exercises

1. Prove that the set of numbers of the form $3n$, where n is an integer, forms a ring.
2. Do the exact decimals (Section 1–10) form a group under addition? A ring?
3. Indicate which of the sets of numbers in Exercise 1, Section 1–14, form rings.
4. Which of the fourteen sets of numbers in Exercise 11, Section 1–17, form rings?
5. Prove that $R[\sqrt{3}] = R(\sqrt{3})$.
6. Write up an intuitive proof of the fact that the complex numbers are algebraically closed.

CHAPTER 2

THEORY OF NUMBERS

The properties of the integers — positive, negative, and zero — may be used to solve several types of problems. They are also of considerable importance in mathematical theories. Thus we follow our development of the number system with a consideration of some of the special properties of the set of integers. Under the definitions and postulates in Chapter 1, the integers

(i) form a commutative group under addition (Section 1–14),
(ii) are closed under multiplication (Section 1–5),
(iii) satisfy the associative law of multiplication (Section 1–5), and
(iv) satisfy the distributive law of multiplication with respect to addition (Section 1–5).

These four properties are precisely the conditions under which a set of elements forms a ring (Section 1–18). Accordingly, we shall speak of the set of integers as the *ring of integers.*

The ring of integers also has three other properties (Sections 1–5 and 1–14) that are not required of all rings:

(i) multiplication is commutative;
(ii) there is an identity for multiplication, unity, in the ring; and
(iii) there are no zero divisors in the ring.

Technically, these additional properties indicate that the ring of integers is also an *integral domain.*

Throughout this chapter we shall be primarily concerned with the properties of the ring of integers. Many of these properties are also properties of more general rings and will be considered as properties of the ring of polynomials in Chapter 3.

2–1 Divisibility. In a field such as the rational number system or the real number system, every element distinct from zero in the field divides every other element. In a ring, such as the ring of integers, the divisibility of an element a by an element $b \neq 0$ cannot be assumed. By definition, an integer b *divides* an integer a (written $b|a$) if and only if $a = bc$, where c is an integer. For example, $2|6$, $3|12$,

and $3|15$, but 3 does not divide 8. The fact that for $c = 0$ every integer $b \neq 0$ divides zero should not be confused with the concept of "zero divisor" (Section 1–14) which is used only when both b and c are different from zero. Using the above definition, we have the following theorem for integers a, b, c.

THEOREM 2–1. *If c divides b and b divides a, then c divides a. If c divides a and c divides b, then c divides $a + b$ and also $a - b$.*

We may take as an example of the first part of the theorem $4|12$, $12|36$, and therefore $4|36$. Similarly, as an example of the second part of the theorem, we may take $4|12$, $4|32$, and therefore $4|44$ where $44 = 32 + 12$, and $4|20$ where $20 = 32 - 12$. These properties may now be proved for arbitrary integers a, b, c satisfying the conditions of the theorem.

In the first part of the theorem it is given that c divides b and that b divides a, that is, that there exist integers r and s such that $b = cr$ and $a = bs$. Then, since multiplication is associative, $a = (cr)s = c(rs)$, whence c divides a. In the last part of the theorem there exist integers p and q such that $a = cp$, $b = cq$ whence, using the distributive law for multiplication with respect to addition, $a + b = c(p + q)$ and $a - b = c(p - q)$. This completes the proof of the theorem.

A number e is a *unit* if e divides every element of the set. The units for the ring of integers are $+1$ and -1. However, only $+1$ is unity, the identity for multiplication.

The integer 2 is a common divisor of 12 and 30. The integers -2, 3, -3, 6, and -6 are also common divisors of 12 and 30. However, 6 is the only positive common divisor of 12 and 30 that is divisible by all other common divisors. We therefore call 6 the greatest common divisor of 12 and 30. In general, we have the following definitions.

If c divides a and c divides b, then c is a *common divisor* of a and b. If c is a positive common divisor of a and b, and every other common divisor d of a and b divides c, then c is *the greatest common divisor* (GCD) of a and b. We write $c = (a, b)$. Any integer ec where $c = (a, b)$ and e is a unit is often called a GCD of a and b. We have modified the usual definition and specified the positive GCD as *the* GCD so that the GCD will be uniquely defined and will coincide with the accepted meaning of the symbol (a, b). Then for any finite set of integers $a_1, a_2, \ldots, a_n$ that are not all zero, we may

define a unique positive greatest common divisor $c = (a_1, a_2, \ldots, a_n)$ with the property that all common divisors of the set are divisors of c. Note that c divides c since $c = c \cdot 1$.

If $c = ka$ and $c = mb$, then c is a *common multiple* of a and b. If c is a positive common multiple of a and b and every other common multiple d of a and b is a multiple of c, then c is *the least common multiple* of a and b. We write $c = [a, b]$. Similarly, for any finite set of integers $a_1, a_2, \ldots, a_n$ that are not all zero, we may define a unique positive least common multiple $c = [a_1, a_2, \ldots, a_n]$ with the property that all common multiples of the set are multiples of c.

Two integers a and b are said to be *relatively prime* if all their common divisors are units, that is, $(a, b) = 1$. For example, $(3, 4) = 1$, $(6, 17) = 1$, $(64, 81) = 1$.

The above definitions of familiar terms are included to keep our thinking rigorous and to serve as a basis for future discussion. In the next section we shall consider a less familiar but very fundamental property of the integers.

Exercises

1. Prove that if $a|b$, then $a|bc$ where c is any integer.
2. Prove that if $a|b$, $a|c$, and $a|d$, then $a|(bx + cy - dz)$, where x, y, z are any integers.
3. Prove that if $0 < a < b$, then b does not divide a.
4. Find all positive integers N such that every positive integer $n \leqq \sqrt{N}$ divides N.
5. Prove (a) that the set of even integers forms a ring and (b) that the set of odd integers does not form a ring. Does the set of even integers also form an integral domain?
6. A "perfect" number is often defined as one that is equal to the sum of its positive divisors (excluding itself). Find the first two such numbers.
7. Prove that the sum of the squares of two odd integers cannot be the square of an integer.
8. Do Exercises 1 through 4, Section 1–18.
9. Indicate which of the sets of numbers in Exercise 1, Section 1–14, form integral domains.

2–2 Division Algorithm. This basic property of the set of positive integers is usually stated as a theorem:

If a and b are any two positive integers, there exist integers q and r, $0 \leqq q$, $0 \leqq r < a$ such that $b = qa + r$.

Given two positive integers 1459 and 112, we might use long division and write $1459 = 13 \cdot 112 + 3$. The above theorem merely establishes that this use of long division is a logical consequence of our previous definitions and theorems. Thus the Division Algorithm, like many of our other theorems, serves to establish a common arithmetic process on the basis of the development of our number system in Chapter 1.

A common "proof" of the Division Algorithm states that either $b = qa$, in which case $r = 0$, or $b \neq qa$ and there exists an integer q such that $qa < b < (q + 1)a$. Thus there always exists an integer r, $0 \leqq r < a$, such that $b = qa + r$. Such a "proof" appears reasonable because it employs only familiar properties of our number system, i.e., if b is divided by a using long division, then either $b = qa$ or $b \neq qa$, and there exists an integer q giving a remainder less than a. However, one purpose in studying the number system is to understand thoroughly what basic assumptions or postulates are necessary. Another purpose is to recognize a "proof" as indicating that the desired result can be obtained from the basic assumptions, definitions, and previously proved theorems. In the above "proof" two points have been overlooked. First, given two positive integers a and b, does there always exist an integer N such that $Na > b$? Second, if there exists at least one integer N as above, does there exist a least such integer, i.e., an integer R such that $(R - 1)a \leqq b < Ra$? The first question may be phrased: Do the positive integers satisfy the Postulate of Archimedes? The second: Are the positive integers well ordered?

The *Postulate of Archimedes* states:

Given any two positive integers a and b, there exists an integer N such that $Na > b$.

This property of the integers may be developed from our previous definitions as follows: Since a is a positive integer, either $a = 1$ or $a > 1$. If $a = 1$, then $ab = b$; if $a > 1$, then from Section 1–6 $(a - 1) > 0$, $(a - 1)b > 0$, $ab - b > 0$, and $ab > b$. In either case, $ab + a = a(b + 1) > b$ and there exists an integer $N = b + 1$ such that $Na > b$. Thus we do not need to assume the Postulate of Archimedes as a postulate, since it can be proved as a theorem on the basis of our earlier definitions.

Finally, a set of elements is said to be *well ordered* (Exercise 6, Section 1–6) if every nonempty subset has a first element. When

ordered according to magnitude, the positive integers are well ordered, negative integers are not, positive rational numbers are not, and the set of rational numbers of the form $1/n$ is not.

The proof that the positive integers are well ordered proceeds as follows: Let I be an arbitrary subset of the positive integers. There are three cases, according as I contains only a finite number of elements, contains infinitely many elements but only a finite number of distinct elements, or contains infinitely many distinct elements. If I contains only a finite number of elements, then there exists a least element of I, since the positive integers are linearly ordered, and one can compare the various elements of a finite set. If I contains infinitely many elements but only a finite number of distinct elements, let N be an upper bound for the finite number of distinct elements of I. Then every element of I coincides with one of the numbers $1, 2, 3, \ldots, N$. Let F be the subset of $1, 2, 3, \ldots, N$ that coincides with the distinct elements of I. Then F is a finite set and has a least element that is also a least element of I. For example, let I be the set $1, 3, 5, 7, 1, 3, 7, 1, 3, 7, \ldots$ and $N = 10$. Each element of I coincides with one of the numbers $1, 2, 3, 4, 5, 6, 7, 8, 9, 10$. The set F is $1, 3, 5, 7$, and 1 is a least element of F and I. If I contains infinitely many distinct elements, let N be an element of I. Then divide the set I into two subsets I_1 and I_2, where those elements of I that are less than or equal to N are in I_1 and all other elements are in I_2. The set I_1 is bounded and has a least element, say R. Since R is less than or equal to N, it is less than all elements of I_2 and is a least element of I.

The assumptions underlying the short "proof" of the Division Algorithm have now been examined in detail. It was found that the assumed properties of the positive integers could be directly proved from the assumptions and definitions of Chapter 1.

The Division Algorithm will be very useful to us. From a practical point of view it underlies the representation to the base ten of all numbers in our Hindu-Arabic or decimal notation (Section 2–7). It is also the basis for a procedure, the Euclidean Algorithm, used to find the greatest common divisor of any two integers (Section 2–5) or of any two polynomials in one variable (Section 3–7).

We next consider prime numbers and lay the foundations for a thorough discussion of the divisors or factors of any given integer in preparation for the Unique Factorization Theorem (Theorem 2–8).

EXERCISES

1. Prove that q and r in the relation $b = qa + r$ of the Division Algorithm are unique.

(*Hint:* Suppose $b = q_1a + r_1 = q_2a + r_2$ where $q_1 \geqq q_2$ and show that $-a < r_2 - r_1 < a$.)

2. State and prove the Postulate of Archimedes for any two rational numbers.

3. What properties must a set of elements such as $1, a, x, 5, r, 2, -1, \ldots$ have before one can attempt to prove that its elements satisfy the Postulate of Archimedes?

4. Indicate which of the following sets of elements are well ordered when ordered algebraically: (a) integers greater than 500, (b) integers greater than -100, (c) negative integers, (d) numbers of the form nk where k is a given positive number and n is any positive integer, (e) positive rational numbers, (f) positive irrational numbers, (g) algebraic numbers.

5. Prove that the elements of any finite or denumerably infinite set of elements may be ordered so that the set is well ordered.

6. Use the result of Exercise 5 and describe an ordering of the rational numbers so that the set is well ordered.

7. Prove that every well-ordered set is linearly ordered (Section 1–6).

2–3 Prime numbers. The following classification of the integers according to the integers that they divide, or are divisible by, will greatly facilitate our study. Zero has been defined (Section 1–5) as the identity under addition. The integers $+1$ and -1 are called units (Section 2–1). An integer p that is not zero or a unit is said to be *prime* if its only divisors are $\pm p$ and the units. An integer is called *composite* if it has two or more prime divisors (not necessarily distinct). For example, $6 = 2 \cdot 3$ and $121 = 11^2$ are composite numbers. We shall find that every integer belongs to one of four classes: zero, units, prime numbers, composite numbers. Since zero is neither positive nor negative, this will mean that every positive integer belongs to one of three classes and every positive integer greater than one is either prime or composite. In the following discussion, negative prime numbers are assumed to be expressed in the form ep, where e is the unit -1 and p is a positive prime. Thus only positive prime numbers need be considered.

We now use these definitions in the proofs of several theorems.

THEOREM 2–2. *Every integer greater than one has a positive prime divisor.*

Let m be any given integer greater than one. Then m is prime if and only if its only positive divisors are m and 1. If m is not a prime, it has a positive divisor m_1, where $m_1 \neq m$ and $m_1 \neq 1$. Thus if m is not prime, it may be written as the product of two positive integers, $m = m_1 m_2$, where neither m_1 nor m_2 is a unit. If neither m_1 nor m_2 is prime, then $m = m_{11}m_{12}m_{21}m_{22}$ where no m_{ij} is a unit. If no m_{ij} is a prime, then $m = m_{111}m_{112}m_{121}m_{122}m_{211}m_{212}m_{221}m_{222}$, where no m_{ijk} is a unit. This process will terminate if and only if at some step at least one of the m's is a prime number. We shall now show that for any given positive integer m the process must terminate, i.e., it cannot continue indefinitely. First, we observe that any positive integer m_2 which is not a unit satisfies the order relation $m_2 > 1$ (Section 1–6). Then we also have $m = m_1 m_2 > m_1$ and, in general,

$$m > m_1 > m_{11} > m_{111} > \cdots$$

for as many steps as the process continues. Thus the process terminates if and only if the set of positive integers $m, m_1, m_{11}, m_{111}, \ldots$ is a finite set. However, this set is a subset of the finite set $m, m-1, m-2, \ldots, 3, 2, 1$ and therefore must itself be finite. Thus the above process must terminate after a finite number of steps, and m must have a prime divisor.

We have also proved that any given positive integer m can have only a finite number of positive integral divisors greater than one. Our next theorem indicates which positive integers need to be considered when one seeks the positive divisors of a given integer m.

THEOREM 2–3. *If a positive integer m is composite, it has a positive prime divisor $\leqq I$, where I is the greatest integer whose square is $\leqq m$.*

By Theorem 2–2, any positive integer m greater than one has a positive prime divisor p, that is, $m = pm_1$. Also, if $m \neq p$, then m_1 has a positive prime divisor $\leqq m_1$. If Theorem 2–3 were false, there would exist a number m that was composite and had no positive prime divisor $\leqq I$. In this case, we would have $I < p$, $I < m_1$ or $I + 1 \leqq p$, $I + 1 \leqq m_1$, and $(I+1)^2 \leqq pm_1 = m$, contrary to the assumption that I is the greatest integer whose square is $\leqq m$. Thus Theorem 2–3 must be true (method of indirect proof, Section 1–10).

Before we can use Theorem 2–3 to determine whether or not a given integer m, say 359, is prime, we need some method for determining the primes $\leqq I$ where $I^2 \leqq m < (I+1)^2$. For the case $m = 359$ we need to know the prime numbers $\leqq 18$.

The prime numbers bounded by any finite integer N may be found by a method called the *Sieve of Eratosthenes:* Write down the integers from 1 to N, exclude 1 since it is a unit, counting from 2 strike out every second number thereafter, counting from 3 strike out every third number, and, in general, counting from any remaining integer k which is $\leqq \sqrt{N}$ (Theorem 2–3) strike out every kth integer. For example, the prime numbers bounded by $N = 18$ are 2, 3, 5, 7, 11, 13, 17, and may be found from the array

~~1~~ 2 3 ~~4~~ 5 ~~6~~ 7 ~~8~~ ~~9~~ ~~10~~ 11 ~~12~~ 13 ~~14~~ ~~15~~ ~~16~~ 17 ~~18~~,

in which it was only necessary to exclude the unit and multiples of 2 and 3 since the next remaining integer, 5, has a square greater than 18.

We now may use Theorem 2–3 and determine whether or not 359 is a prime by testing 359 successively for divisibility by 2, 3, 5, 7, 11, 13, 17. On this basis we may assert that 359 is a prime number.

One reason for considering such mechanical methods as the above for determining primes lies in the fact that no analytical representation or formula for all primes has yet been found. However, we may prove several theorems regarding primes. The following theorem is a modern version of Proposition 20 in Book IX of Euclid's *Elements.*

THEOREM 2–4. *The set of positive prime numbers is countably infinite.*

Suppose there were a largest prime number, say P, then the number $N = P! + 1$ must have a prime divisor (Theorem 2–2). But no number $\leqq P$ divides $P! + 1 = N$. Thus N has a prime divisor greater than P and there is no greatest prime, i.e., the set of positive prime numbers is countably infinite. For example, if $P = 2$, then $N = 2! + 1 = 3$, which is prime; if $P = 5$, then $N = 5! + 1 = 121$, which has $11 > 5$ as a prime divisor. This process for determining the existence of a prime greater than a given prime P may also be used, together with the existence of a single prime number 2, to prove by mathematical induction (Section 1–4) that there exists a countably infinite subset of the set of positive prime numbers. Then, since the set of all positive prime numbers is a subset of the set of positive integers, which is countably infinite, we have another proof that the set of positive prime numbers is countably infinite.

The best-known properties of primes concern divisibility. Given any integer m and prime p, the only positive divisors of p, and therefore the only possible positive common divisors of p and m, are p and 1. Thus we have

THEOREM 2–5. *If p is a prime and m is any integer, then either p divides m or $(p, m) = 1$.*

Another common theorem may be proved as follows: Suppose p is a prime number, and a and b are each positive integers less than p. We wish to prove that p does not divide the product ab, written $p \nmid ab$. We shall use the method of indirect proof and suppose that $p|ab$. Furthermore, we shall assume that b is the smallest positive integer such that $p|ab$, that is, ab is the least multiple of a such that $p|ab$. This last assumption may be made without loss of generality, since if there exists a single integral multiple, there must be a smallest positive integral multiple of a that is divisible by p (Section 2–2). Now by the Division Algorithm, there exists an integer m such that

$$mb \leqq p < (m+1)b, \quad 0 \leqq p - mb < b.$$

Actually $mb \neq p$ since $1 < b < p$ and p is prime. By assumption, $p|ab$ and thus $p|mab$. Then from $p|ap$ we have $p|(ap - mab)$ and $p|a(p - mb)$, whence $a(p - mb)$ is a multiple of a which is divisible by p. But also $a(p - mb) < ab$, contrary to the assumption that ab is the least multiple of a that is divisible by p. Thus p does not divide ab, and we have given an indirect proof of the following theorem.

THEOREM 2–6. *If p is a prime, and a and b are two positive integers each less than p, then $p \nmid ab$.*

This theorem may be extended to include any two positive integers a and b such that $p \nmid a$ and $p \nmid b$. Let $a = mp + r$, $b = np + s$, $0 < r < p$, $0 < s < p$. Now if $p|ab$, we also have $p|rs$, contrary to Theorem 2–6. Thus if $p \nmid a$ and $p \nmid b$, then $p \nmid ab$. In other words, if $p|ab$, then $p|a$ or $p|b$. Since the product of two integers is an integer, we may also take $a_1 \cdot a_2 = a$, $a_3 = b$ and prove that if $p|a_1a_2a_3$, then p divides at least one of the numbers a_1, a_2, a_3. By repeated application of this process, we have

THEOREM 2–7. *If p is a prime and $p|a_1a_2 \ldots a_n$, then p divides at least one of the integers $a_1, a_2, \ldots, a_n$, where n is any positive integer.*

A very important application of this property of prime numbers is found in the factorization of all positive integers as products of powers of prime numbers (Section 2–4). Throughout the remainder of this text we shall make extensive use of the properties of prime numbers and the analogous properties of irreducible polynomials (Section 3–6).

EXERCISES

1. Find the prime numbers less than 200, using the Sieve of Eratosthenes.
2. Determine which of the following are prime numbers:
   (a) 85, 103, 179, 539,
   (b) 267, 781, 859, 937,
   (c) 1245, 2287.
3. Write out a formal proof of Theorem 2–7, using mathematical induction.
4. Is $n^2 - n + 41$ a prime number for all positive integral values of n? Explain.
5. Give four numerical examples illustrating Theorem 2–5.
6. Repeat Exercise 5 for Theorems 2–6 and 2–7.
7. Given any integer N, how could you find *all* its positive prime divisors?
8. Prove that $n^3 + 1$ is a composite number if n is greater than one.
9. Prove that $3^n - 1$ and, in general, $m^n - 1$ is composite if n is greater than one and m is greater than 2 (see Exercise 7, Section 1–4).
10. A number of the form $2^p - 1$ that is prime is called a *Mersenne prime*. Find five such numbers.
11. Prove that $2^n - 1$ is composite if n is composite (see Exercise 9, Section 1–4). Give an example of a composite number of the form $2^p - 1$ where p is a prime.

2–4 Unique Factorization Theorem. An integer is said to be completely factored when it is expressed as a product of prime numbers (assumed positive) and a unit (+1 or −1). In this section we shall first use the fact that the product of any finite number of units is a unit and show that any integer may be expressed as a product of positive prime numbers and a unit in essentially only one way. Then we shall draw some further results from this factorization.

The positive integer 168 may be expressed as a product of integers in several ways. For example,

$$168 = 4 \cdot 42 = 2 \cdot (-2) \cdot (-7) \cdot 6 = 21 \cdot 8 = 7 \cdot 24.$$

The unique factorization theorem states that when 168 is expressed as a product of positive prime numbers, $168 = 2^3 \cdot 3 \cdot 7$, any other

factorization such as $168 = 3 \cdot 2^3 \cdot 7$ into prime divisors must coincide with the first one, except possibly for the order in which the divisors are written.

Any positive integer m greater than 1 has at least one positive prime divisor or factor, by Theorem 2–2. Such a prime divisor, say p_1, can be found in a finite number of steps, since m is finite and p_1 is one of the numbers $1, 2, 3, \ldots, m$. If $p_1 = m$, our factorization is complete and unique. If $p_1 \neq m$, let $m = p_1 m_1$ and proceed as before with m_1, obtaining $m = p_1(p_2 m_2)$ if m_1 is not prime. Since the positive integers $m, m_1, m_2, \ldots$ satisfy $m > m_1 > m_2 > \ldots$, the above process, like the one in the proof of Theorem 2–2, must terminate after a finite number of steps and give

$$m = p_1 p_2 \ldots p_r. \tag{2–1}$$

If there were also a second factorization,

$$m = q_1 q_2 \ldots q_s \tag{2–2}$$

of m into positive prime divisors, we would have

$$p_1 p_2 \ldots p_r = q_1 q_2 \ldots q_s. \tag{2–3}$$

Since p_1 divides $q_1 q_2 \ldots q_s$, it must, by Theorem 2–7, divide some q_i, say q_1. Since q_1 and p_1 are assumed to be positive prime numbers, $p_1 = q_1$. We divide both sides of (2–3) by $p_1 = q_1$ and repeat the same argument to show that p_2 is equal to one of the q's, say q_2. This process may be continued until one side of (2–3) is reduced to 1. Since the p's and q's are integers, the other side must simultaneously become 1. Thus there exists a factorization of m into prime divisors, and any two factorizations (2–1) and (2–2) of m into prime divisors are identical, except possibly for the order in which the divisors are written. The divisors and thus the factorization are unique. If the equal primes are grouped together, we have the *Unique Factorization Theorem* or, as it is sometimes called, the *Fundamental Theorem of Arithmetic:*

THEOREM 2–8. *Every integer except zero can be represented in one and only one way in the form*

$$m = e_i p_1^{a_1} p_2^{a_2} \ldots p_n^{a_n},$$

where e_i is one of the units, the p_i are distinct positive primes, and the a_i are positive integers.

Given any integer m, we may therefore select the appropriate unit by observation and then use Theorem 2–3 and successive division to find the positive prime divisors of m. For example, $12 = 2^2 \cdot 3$, $-36 = (-1) \cdot 2^2 \cdot 3^2$, $1232 = 2^4 \cdot 7 \cdot 11$.

Let us consider $12 = 2^2 \cdot 3$ for a moment. Any prime divisor of 12 must divide 2^2 or 3, by Theorem 2–7. Thus 2 and 3 are the only prime divisors of 12. Similarly, all positive divisors d of 12 may be expressed in the form $d = 2^a \cdot 3^b$, where $a = 0, 1, 2$ and $b = 0$ or 1. All of the divisors of $12 = 2^2 \cdot 3$ are of the form $e \cdot 2^a \cdot 3^b$, where e is a unit, $0 \leqq a \leqq 2$, and $0 \leqq b \leqq 1$. In general, all the divisors of $m = e_i p_1^{a_1} p_2^{a_2} \ldots p_n^{a_n}$ are of the form

$$e_k p_1^{b_1} p_2^{b_2} \ldots p_n^{b_n}, \text{ where } 0 \leqq b_i \leqq a_i \tag{2–4}$$

and e_k is a unit. Furthermore, every number of the form (2–4) is a divisor of m. From this concept of divisor and Theorem 2–8, we have

THEOREM 2–9. *If a and b have no divisors in common and each divides c, then their product divides c. If a and c have no common divisors and b and c have no common divisors, then ab and c have no common divisors. If a and c have no common divisors and c divides ab, then c divides b.*

The three parts of this theorem may be stated in mathematical notation as follows: (i) $(a, b) = 1$, $a|c$ and $b|c$ imply $ab|c$; (ii) $(a, c) = 1$ and $(b, c) = 1$ imply $(ab, c) = 1$; (iii) $(a, c) = 1$ and $c|ab$ imply $c|b$. The proofs of these statements are given as exercises (Exercises 3, 4, and 5).

Theorem 2–8 may also be used to find the greatest common divisor and lowest common multiple of two integers (Section 2–1). For example, if $m = 2^3 \cdot 3^2 \cdot 5^2 \cdot 7$ and $n = 2^2 \cdot 3^3 \cdot 7^2 \cdot 11$, then $(m, n) = 2^2 \cdot 3^2 \cdot 7$ and $[m, n] = 2^3 \cdot 3^3 \cdot 5^2 \cdot 7^2 \cdot 11$. These particular values may be obtained by observation. In general, it is often advantageous to allow zero exponents in order to express both m and n in terms of the same positive prime numbers. For example, in the above case, $m = 2^3 \cdot 3^2 \cdot 5^2 \cdot 7 \cdot 11^0$ and $n = 2^2 \cdot 3^3 \cdot 5^0 \cdot 7^2 \cdot 11$. Thus, given any integers m and n, we may write each in terms of its positive prime divisors and then rewrite each as above in terms of the same set of prime numbers, say $m = e_i p_1^{a_1} p_2^{a_2} \ldots p_k^{a_k}$ and $n = e_r p_1^{b_1} p_2^{b_2} \cdots p_k^{b_k}$. These expressions may be abbreviated, using the product symbol $\prod$, as

$$m = e_i \prod_{i=1}^{k} p_i^{a_i}, \quad n = e_r \prod_{i=1}^{k} p_i^{b_i}.$$

Then (m, n) is obtained by taking the smallest occurring exponent on each prime number p_i, and $[m, n]$ is obtained by taking the largest occurring exponent on each prime number p_i. In mathematical notation, we write

$$(m, n) = \prod_{i=1}^{k} p_i^{c_i}, \quad [m, n] = \prod_{i=1}^{k} p_i^{d_i},$$

where c_i is the minimum of a_i and b_i, d_i is the maximum of a_i and b_i.

Finally, suppose $(a, b) = d$, $[a, b] = m$, and let $a = a_1 d$, $b = b_1 d$. Then $(a_1, b_1) = 1$, the lowest common multiple of a and b is $m = a_1 b_1 d$, and $dm = da_1 b_1 d = ab$. Thus we have

THEOREM 2–10. *If a and b are positive integers, $(a, b) = d$, and $[a, b] = m$, then $dm = ab$.*

For example, $(6, 8) = 2$, $[6, 8] = 24$, and $6 \cdot 8 = 2 \cdot 24$. The fact that this theorem cannot be extended directly to the case of three positive integers is evident from the following example: $(6, 4, 10) = 2$, $[6, 4, 10] = 60$, and $6 \cdot 4 \cdot 10 = 240 \neq 2 \cdot 60$.

In the next section we shall use the Euclidean Algorithm to find $d = (a, b)$ without first expressing a and b in terms of their prime divisors. Then we may use $m = ab/d$ from Theorem 2–10 to find $m = [a, b]$. Thus, we shall soon be able to find both (a, b) and $[a, b]$ without expressing a and b in terms of their prime divisors.

EXERCISES

1. Factor the numbers 4680, 1275, and 1273 in terms of their positive prime divisors.

2. Find (4680, 1275) and [4680, 1275] in terms of their prime divisors. Does Theorem 2–10 hold?

3. Prove the first part of Theorem 2–9.

4. Prove the second part of Theorem 2–9.

5. Prove the third part of Theorem 2–9.

6. Given any integer n, how could you find *all* its positive divisors?

7. Find all the positive divisors of 60.

8. Prove that every positive divisor of $m = e_i p_1^{a_1} p_2^{a_2} \cdots p_n^{a_n}$ occurs once and only once among the terms of the product

$$(1 + p_1 + p_1^2 + \cdots + p_1^{a_1})(1 + p_2 + p_2^2 + \cdots + p_2^{a_2}) \cdots (1 + p_n + p_n^2 + \cdots + p_n^{a_n}).$$

9. Prove that the integer m in Exercise 8 has $(a_1 + 1)(a_2 + 1) \cdots a_n + 1)$ distinct positive divisors.

10. Prove that the sum of the positive divisors of the integer m in Exercise 8 may be expressed in the form

$$\prod_{i=1}^{n} \frac{p_i^{a_i+1} - 1}{p_i - 1},$$

using the product symbol $\prod$.

11. Find the number and sum of the positive divisors of 60, using Exercises 9 and 10.

12. How many divisors has each of the numbers in Exercise 1?

13. Find the sum of the divisors of each of the numbers in Exercise 1.

2–5 Euclidean Algorithm. The greatest common divisor (m, n) of two integers was expressed in Section 2–4 in terms of the prime divisors of the two integers. The Euclidean Algorithm gives a straightforward method for obtaining the greatest common divisor of two integers without expressing either integer in terms of its prime divisors. This method is especially advantageous when large numbers are involved. In the case of 36 and 90, the method of Section 2–4 would be to write $36 = 2^2 \cdot 3^2$ and $90 = 2 \cdot 3^2 \cdot 5$, then $(36, 90) = 2 \cdot 3^2 = 18$. The Euclidean Algorithm would give $90 = 2 \cdot 36 + 18$, $36 = 2 \cdot 18 + 0$ and $(36, 90) = 18$.

In general, since the greatest common divisor is taken as positive, it may be obtained for any two integers different from zero by considering only the corresponding positive integers m, n where factors of $+1$ or -1 have been inserted. If $m = n$, then also $(m, n) = m$; if $m \neq n$, we may suppose $m > n$. Then we repeatedly apply the Division Algorithm (Section 2–2) and obtain the *Euclidean Algorithm*:

$$\begin{aligned}
&(2\text{–}5) && m = qn + n_1, && 0 < n_1 < n \\
&(2\text{–}6) && n = q_1 n_1 + n_2, && 0 < n_2 < n_1 \\
&(2\text{–}7) && n_1 = q_2 n_2 + n_3, && 0 < n_3 < n_2 \\
& && \cdot && \cdot \\
& && \cdot && \cdot \\
& && \cdot && \cdot \\
&(2\text{–}8) && n_{k-2} = q_{k-1} n_{k-1} + n_k, && 0 < n_k < n_{k-1} \\
&(2\text{–}9) && n_{k-1} = q_k n_k. && 0 = n_{k+1}
\end{aligned}$$

Since the integers $n, n_1, n_2, \ldots$ form a decreasing sequence, i.e., $n > n_1 > n_2 > \ldots$, there exists some n_j, say n_{k+1}, equal to zero and such that either $k = 0$ or n_k is different from zero. We shall find that $(m, n) = n = n_0$ when $k = 0$, and $(m, n) = n_k$ when $k \neq 0$.

Any common divisor of m and n must divide n_1 by (2–5), n_2 by (2–6), n_3 by (2–7), . . . , n_k by (2–8). Thus, every common divisor of m and n is a divisor of n_k. Conversely, n_k divides n_{k-1} by (2–9), n_{k-2} by (2–8), . . . , n_1 by (2–7), n by (2–6), and m by (2–5), that is, n_k is a common divisor of m and n. These results are stated in the following theorem:

THEOREM 2–11. *The greatest common divisor of any two positive integers m and n can be found as the last nonvanishing remainder in the Euclidean Algorithm. There exist integers A and B such that*

$$(m, n) = n_k = Am + Bn. \tag{2–10}$$

The integers A and B in (2–10) may be obtained by solving (2–5) for n_1 in the form $n_1 = A_1m + B_1n$ and substituting this in (2–6) to obtain $n_2 = A_2m + B_2n, \ldots$; finally $n_k = Am + Bn$ from (2–8). For example, we have the following form of the Euclidean Algorithm for the numbers 23 and 19:

$$\begin{aligned} 23 &= 1 \cdot 19 + 4, \\ 19 &= 4 \cdot 4 + 3, \\ 4 &= 1 \cdot 3 + 1, \\ 3 &= 3 \cdot 1 + 0, \end{aligned}$$

whence (23, 19) = 1. A relation of the form (2–10) may be found as indicated above, using the equations

$$\begin{aligned} 4 &= 1 \cdot 23 - 1 \cdot 19, \\ 3 &= 1 \cdot 19 - 4 \cdot 4 = 5 \cdot 19 - 4 \cdot 23, \\ 1 &= 1 \cdot 4 - 1 \cdot 3 = 5 \cdot 23 - 6 \cdot 19. \end{aligned}$$

This process may be expressed in terms of the general quotients q_i and remainders n_i in (2–5) to (2–9) as follows:

$$\begin{aligned} n_1 &= m - qn = A_1m + B_1n, \\ n_2 &= -q_1m + (q_1q + 1)n = A_2m + B_2n, \\ n_3 &= (q_1q_2 + 1)m - (q_2 + qq_1q_2 + q)n = A_3m + B_3n, \\ &\cdot \\ &\cdot \\ &\cdot \end{aligned}$$

Since only the ring operations of addition, subtraction, and multiplication are involved, the coefficients of m and n are integers at every step.

There still remains the practical problem of finding A and B for any given integers m and n with as little work as possible. For any

integer $m \neq 0$, we have $(m, 0) = em$ where e is a unit. Also, we have observed that the given integers may be assumed positive without any loss of generality. The following array gives a practical procedure for determining both the greatest common divisor and a relation of the form (2–10) for any two positive integers m and n. The numbers n_i and q_i are the same as in (2–5) to (2–9). The array

$$\begin{array}{cccccccc} m & n & n_1 & n_2 & n_3 & \ldots & n_k & 0 \\ & q & q_1 & q_2 & q_3 & \ldots & q_k & \\ & 1 & -q & B_2 & B_3 & \ldots & B_k = B & \\ & & 1 & -q_1 & A_3 & \ldots & A_k = A & \end{array} \tag{2–11}$$

may be constructed by using the geometrical pattern

$$\begin{array}{ccc} n_{i-1} & n_i & n_{i+1} \\ & q_i & \end{array}$$

in the first two rows to signify that $n_{i-1} = q_in_i + n_{i+1}$, representing the last n_j different from zero by n_k, and determining the A_j and B_j by the recurrence formulas

$$\begin{aligned} &B_0 = 1, A_0 = 0, B_1 = -q, A_1 = 1, \\ &B_{i+1} = B_{i-1} - q_iB_i, \\ &A_{i+1} = A_{i-1} - q_iA_i. \end{aligned} \qquad (i = 1, 2, 3, 4, \ldots, k - 1)$$

In the case of $m = 23$, $n = 19$, this array becomes

$$\begin{array}{cccccc} 23 & 19 & 4 & 3 & 1 & 0 \\ & 1 & 4 & 1 & 3 & \\ & 1 & -1 & 5 & -6 & \\ & & 1 & -4 & 5 & \end{array} \tag{2–12}$$

whence (23, 19) = 1 and $1 = 5 \cdot 23 - 6 \cdot 19$, as was previously obtained.

The above method of obtaining n_k is merely a synthetic representation of the relations (2–5) to (2–9) and is therefore valid. The above method of determining A_k and B_k may be verified by using mathematical induction upon j, where $n_j = mA_j + nB_j$. For $j = 0$ we take $n_0 = n$ and have $n = n$; for $j = 1$ we have $n_1 = m - qn$, which is valid from (2–5). Suppose $n_{i-1} = mA_{i-1} + nB_{i-1}$ and $n_i = mA_i + nB_i$, then

$$\begin{aligned} n_{i+1} &= n_{i-1} - q_in_i \\ &= m(A_{i-1} - q_iA_i) + n(B_{i-1} - q_iB_i) \\ &= mA_{i+1} + nB_{i+1}, \end{aligned}$$

verifying the recurrence formulas given above.

The method represented by the array (2–11) can be very useful after a little practice. It is especially advantageous in that the greatest common divisor can be determined by using only the first two rows. Then if a relation of the form (2–10) is desired, the constants A and B can be determined.

The equation (2–10) is now both necessary and sufficient for n_k to be the greatest common divisor of m and n. It is necessary by Theorem 2–11 and sufficient since if (2–10) holds, every common factor of m and n must divide n_k. As a particular application of this, we have

THEOREM 2–12. *Two integers m and n are relatively prime if and only if there exist integers A and B such that $Am + Bn = 1$.*

We shall use the Euclidean Algorithm and Theorem 2–12 to obtain the reciprocal of n modulo m in Section 2–11. In Section 3–7 both of these results are restated for polynomials $p(x)$. In this form they will be used to find the greatest common divisor of two polynomials, the number of distinct roots of a polynomial equation on any interval $a < x \leqq b$ (Section 4–12), and the multiple roots of a polynomial equation (Section 4–13).

EXERCISES

1. How may the Euclidean Algorithm be used to prove the existence of a greatest common divisor for any two positive integers?

2. Prove that $(km, kn) = k(m, n)$ for any positive integer k.

3. Express each of the following in the form of (2–10):

(a) (108, 64),
(b) (370, 111),
(c) (147, 64),
(d) (3961, 952),
(e) (4680, 1275).

Compare with Exercise 2, Section 2–4.

4. Find the lowest common multiple of each of the pairs of numbers used in Exercise 3.

5. Prove that $[km, kn] = k[m, n]$.

6. Prove that if $(a, b) = 1$, where a and b are any integers, then there exist integers d and e such that

$$\frac{1}{ab} = \frac{d}{a} + \frac{e}{b}.$$

7. Prove that every positive rational number may be expressed as a terminating continued fraction of the form

$$\frac{m}{n} = q + \cfrac{1}{q_1 + \cfrac{1}{q_2 + \cfrac{1}{q_3 + \cfrac{1}{\ddots + \cfrac{1}{q_k}}}}}$$

which is frequently written as

$$\frac{m}{n} = q + \frac{1}{q_1^+}\ \frac{1}{q_2^+}\ \frac{1}{q_3^+} \cdots + \frac{1}{q_k}.$$

(*Hint:* The q's here are the same as in the Euclidean Algorithm. Corresponding to (2–12), we have $\frac{23}{19} = 1 + \frac{1}{4^+}\ \frac{1}{1^+}\ \frac{1}{3}$.)

8. Use the results obtained in Exercise 3 and express each of the following rational numbers as continued fractions: $\frac{108}{64}, \frac{370}{111}, \frac{147}{64}, \frac{3961}{952}, \frac{4680}{1275}$.

9. Prove that $-B/A$ in (2–10) is an approximation for m/n differing from m/n by n_k/An. This is a very practical first approximation and is equivalent to dropping the term q_k in Exercise 7. Thus $\frac{23}{19}$ is approximately $\frac{6}{5}$, and in this case the error is $\frac{1}{114}$.

10. Use the method of Exercise 9 and find first approximations to each of the fractions in Exercise 8. Indicate the error of the approximation in each case.

11. Continue the process of approximation started in Exercise 9 and show that in general the $(j + 1)$th approximation of m/n is $-B_{k-j}/A_{k-j}$ in (2–11).

2–6 Bases. The concept of a "base" is as fundamental in theory of numbers as it is in baseball. Any representation of numbers, such as 1776, in which the position of the digits has significance depends upon the particular number that is used as a base. For example, 11 represents eleven to the base ten, i.e., one ten and one unit in the decimal notation. However, 11 also represents three to the base two, i.e., one two and one unit in the binary number system. We are all familiar with the decimal notation (Section 2–7) using the base ten and the digits 0, 1, 2, 3, 4, 5, 6, 7, 8, 9. The binary system

uses only two digits, say 0 and 1, and is becoming of increasing importance with the development of electronic computers, since its digits may be represented by the presence and absence, respectively, of an electric current. In general, we shall prove that every positive integer n greater than 1 can be used as a base for all positive integers, that is,

THEOREM 2–13. *If m and n are positive integers, $n > 1$, then the representation*

$$m = a_k n^{k-1} + a_{k-1} n^{k-2} + \cdots + a_1$$

where $a_k \neq 0$, $0 \leqq a_i < n$ for $i = 1, 2, \ldots, k$, exists for some integer k and is unique.

The integer 130, for example, may be expressed using bases 10, 2, 3, 4, 5, and 6 as follows:

$$130 = 1 \cdot 10^2 + 3 \cdot 10 + 0 = 130_{10},$$

(2–13)
$$\begin{aligned} 130 &= 1 \cdot 2^7 + 0 \cdot 2^6 + 0 \cdot 2^5 + 0 \cdot 2^4, \\ &\quad + 0 \cdot 2^3 + 0 \cdot 2^2 + 1 \cdot 2 + 0 = 10000010_2, \\ 130 &= 1 \cdot 3^4 + 1 \cdot 3^3 + 2 \cdot 3^2 + 1 \cdot 3 + 1 = 11211_3, \\ 130 &= 2 \cdot 4^3 + 0 \cdot 4^2 + 0 \cdot 4 + 2 = 2002_4, \\ 130 &= 1 \cdot 5^3 + 0 \cdot 5^2 + 1 \cdot 5 + 0 = 1010_5, \end{aligned}$$

(2–14)
$$130 = 3 \cdot 6^2 + 3 \cdot 6 + 4 = 334_6.$$

The integers a_i may be found by means of repeated applications of the Division Algorithm (Section 2–2), but the process is quite different from that used in Section 2–5. For example, in (2–13) and (2–14), we have

$$\begin{aligned} 130 &= 10 \cdot 13 + 0, & 130 &= 6 \cdot 21 + 4, \\ 13 &= 10 \cdot 1 + 3, & 21 &= 6 \cdot 3 + 3, \\ 1 &= 10 \cdot 0 + 1, & 3 &= 6 \cdot 0 + 3. \end{aligned}$$

The successive remainders 0, 3, 1 when 130 is repeatedly divided as above by 10 are the units, tens, and hundreds digits, i.e., the coefficients of 1, 10, and 10^2 in (2–13). Similarly, the successive remainders 4, 3, 3 when 130 is repeatedly divided as above by 6 are the coefficients of 1, 6, and 6^2 in (2–14). These coefficients are easily obtained using the following arrays, where the remainders are set off to the side:

$$\begin{array}{r|l} 10 & 130 \\ \hline 10 & 13 \sim 0 \\ 10 & 1 \sim 3 \\ & 0 \sim 1 \end{array} \qquad \begin{array}{r|l} 6 & 130 \\ \hline 6 & 21 \sim 4 \\ 6 & 3 \sim 3 \\ & 0 \sim 3 \end{array}$$

In general, the coefficients a_i in Theorem 2–13 may be obtained as follows. Suppose

$$\begin{aligned} m &= q_1 n + r_1, \\ q_1 &= q_2 n + r_2, \\ q_2 &= q_3 n + r_3, \\ &\cdots \\ q_{k-2} &= q_{k-1} n + r_{k-1}, \\ q_{k-1} &= 0 \cdot n + r_k, \end{aligned}$$

where $0 \leqq r_i < n$. From these equations, we have $m > q_1 > q_2 > \cdots > q_{k-1} > 0$, $0 < q_{k-1} < n$, $r_k = q_{k-1}$, and therefore $0 < r_k$. The coefficients in Theorem 2–13 are obtained by taking $a_k = r_k$, $a_{k-1} = r_{k-1}, \ldots, a_1 = r_1$.

The relations $a_j = r_j$ may be verified by substituting each q_i in the equation $q_{i-1} = q_i n + r_i$ for $i = k-1, k-2, \ldots, 2, 1$, and $q_0 = m$, as follows:

$$\begin{aligned} q_{k-1} &= r_k, \\ q_{k-2} &= r_k n + r_{k-1}, \\ q_{k-3} &= r_k n^2 + r_{k-1} n + r_{k-2}, \\ &\cdots \\ q_1 &= r_k n^{k-2} + r_{k-1} n^{k-3} + \cdots + r_2, \\ m &= r_k n^{k-1} + r_{k-1} n^{k-2} + \cdots + r_1. \end{aligned}$$

If there were two expressions for m in terms of the same base n, say

$$m = a_0 n^k + a_1 n^{k-1} + \cdots + a_{k-1} n + a_k, \quad 0 < a_0, 0 \leqq a_k < n$$

and

$$m = b_0 n^p + b_1 n^{p-1} + \cdots + b_{p-1} n + b_p, \quad 0 < b_0, 0 \leqq b_i < n,$$

then we would have

$$a_0 n^k + a_1 n^{k-1} + \cdots + a_k - (b_0 n^p + b_1 n^{p-1} + \cdots + b_p) = 0.$$

The right member, 0 (and therefore the left member of this equation) is divisible by n. Thus n divides $a_k - b_p$. But

$$0 \leqq a_k < n, 0 \leqq b_p < n, |a_k - b_p| < n,$$

and n does not divide any positive integer less than n. Since $|a_k - b_p|$ is divisible by n, it cannot be positive and must be equal to zero,

that is, $a_k = b_p$. Similarly, both members must be divisible by n^{j+1}, whence $a_{k-j} = b_{p-j}$ or, in general, $a_i = b_i$ for all values of i, and we have $p = k$. Thus the representation of m in Theorem 2–13 is unique.

The Hindu-Arabic notation or decimal system (Section 2–7) which we use consists of numbers expressed to the base 10. In this case Theorem 2–13 states that every positive integer has a unique representation to the base 10. For example,

$$5604 = 5 \cdot 10^3 + 6 \cdot 10^2 + 0 \cdot 10 + 4.$$

Similarly, for $n = 2$ Theorem 2–13 states that every positive integer has a unique representation to the base 2. For example,

$$183 = 1 \cdot 2^7 + 0 \cdot 2^6 + 1 \cdot 2^5 + 1 \cdot 2^4 + 0 \cdot 2^3 + 1 \cdot 2^2 + 1 \cdot 2 + 1,$$

and may be indicated by 10110111_2. The binary system may be extended in the same manner as the decimal system to represent $7.625 = 1 \cdot 2^2 + 1 \cdot 2 + 1 \cdot 2^0 + 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3}$ by 111.101_2. The binary system forms the basis for Russian peasant multiplication, an ancient method of reducing the multiplication of two integers to addition (Exercises 11 and 12). More recently, as mentioned above, the binary system has been used as the basis for calculations with electronic computers.

There are many problems and games depending upon the scale of notation, i.e., the base to which the numbers are expressed. Ball [3; 11–16] indicates several problems in which the base ten is used. For example, if your friend selects two integers less than 10 (possible by a throw of two dice), you may discover the integers by asking him to

(i) select one of the integers and multiply it by 5,
(ii) add 7,
(iii) multiply by 2,
(iv) add in the second integer, and tell you the result.

From the following algebraic considerations of the above procedure

(i) $5a$,
(ii) $5a + 7$,
(iii) $10a + 14$,
(iv) $10a + 14 + b$,

it is clear that you need only to subtract 14 from the number given by your friend to obtain a number whose two digits are precisely his original integers.

The game of Nim [50; 16–19] can be completely analyzed by means of the binary numbers. Another game involving binary numbers [43; 39] requires k cards where numbers less than 2^k will be considered. All positive integers less than 2^k have a unique expansion as a sum of powers of 2 in the form

$$n = a_1 + a_2 \cdot 2 + a_3 \cdot 2^2 + \cdots + a_k \cdot 2^{k-1}, \tag{2–15}$$

where $a_i = 0$ or 1 $(i = 1, 2, \ldots, k)$. The first card contains all positive integers less than 2^k for which $a_1 = 1$, the second those for which $a_2 = 1, \ldots,$ the kth those for which $a^k = 1$. The first number on the jth card is 2^{j-1}. Thus in order to determine a number it is only necessary to know on which cards it occurs, i.e., the powers of 2 that are used in its binary expansion. Anyone familiar with the game can then state the number without looking at the cards, since its representation as a binary number has been given. The desired number is the sum of the first numbers on each of the cards on which it occurs. For example, if $k = 4$, we have cards

1	9	2	10	4	12	8	12
3	11	3	11	5	13	9	13
5	13	6	14	6	14	10	14
7	15	7	15	7	15	11	15

The number $5 = 1 + 0 \cdot 2 + 1 \cdot 2^2$ occurs only on the first and third cards. The number that occurs only on the second and third cards is $0 \cdot 2^0 + 1 \cdot 2 + 1 \cdot 2^2 = 6$. The number that occurs only on the first, third, and fourth cards is $1 \cdot 2^0 + 0 \cdot 2 + 1 \cdot 2^2 + 1 \cdot 2^3 = 13$. In general, the binary representation of the number is known as soon as the cards on which the number occurs are known. In (2–15) $a_1 = 1$ if the number n is on the first card, $a_1 = 0$ if n is not on the first card. Similarly, considering $n < 16$ or a larger set of cards than those given above, $a_i = 1$ in (2–15) if n is on the ith card, $a_i = 0$ otherwise.

Exercises

1. Express 19 and 175 to the base 2.
2. Express 95 and 348 to the base 3.
3. Express 75 and 6789 to the base 7.
4. In each of the above exercises, add the two numbers, using the new

base. Check this addition by adding the two numbers as given to the base ten and expressing the sum to the indicated base.

5. Repeat Exercise 4, using subtraction of the first number from the second in place of addition.

6. Repeat Exercise 4, using multiplication.

7. Repeat Exercise 4, using division of the second number by the first.

8. Express the number 12143_5 (where the subscript indicates the base) to the base 10 and then to the base 7.

9. Express 12143_5 to the base 7 without changing it to the base 10.

10. Give a general method for changing the base of any integer. Illustrate your method using three integers of at least five digits each where all bases are different from ten.

11. The following is an example of Russian peasant multiplication of $43 \cdot 75$. The integral parts of the successive quotients of 43 by 2 are listed beside the successive multiples of 75 by 2. Then those multiples of 75 corresponding to odd quotients of 43 are added to obtain the desired product.

43	75
21	150
(10)	(300)
5	600
(2)	(1200)
1	2400

$43 \cdot 75 = 75 + 150 + 600 + 2400 = 3225$. Use a binary representation and give a proof of the validity of this method.

12. Find the following products by Russian peasant multiplication: $67 \cdot 85$, $73 \cdot 120$, $121 \cdot 373$.

13. The cancellation $\frac{16}{64}$ gives a correct answer for the fraction $\frac{16}{64}$ to the base ten. Use the base ten and find all fractions m/n, where $10 < m < 20$, $10 < n < 100$, such that similar cancellations give a correct answer.

14. Find all fractions m/n such that m, n are two digit numbers to the base ten and a cancellation similar to that used in Exercise 13 gives a correct answer.

15. Prove that there are no fractions such as those sought in Exercise 14 when the numbers are expressed to the base p, where p is a prime number.

2–7 Decimal notation. The representation of a number to the base 10 is called the *decimal notation* for the number. We have found in Section 2–6 that any positive integer m has a unique representation in decimal notation, i.e.,

$$m = a_k 10^k + a_{k-1} 10^{k-1} + \cdots + a_1 10 + a_0,$$

where

$$\begin{aligned} m &= 10m_1 + a_0, \\ m_1 &= 10m_2 + a_1, \\ &\cdot \\ &\cdot \\ &\cdot \\ m_k &= 10 \cdot 0 + a_k, \end{aligned}$$

and $0 \leqq a_j < 10$, $a_k \neq 0$.

We now consider the representation in decimal notation of positive rational numbers s/n. This representation is also based upon the Division Algorithm. For example, given the number $\frac{51}{8}$, we may consider

$$\begin{aligned} 51 &= 6 \cdot 8 + 3, \\ 30 &= 3 \cdot 8 + 6, \\ 60 &= 7 \cdot 8 + 4, \\ 40 &= 5 \cdot 8 + 0, \end{aligned}$$

and write $\frac{51}{8} = 6.3750$. In general, given any positive rational number s/n we may use the Division Algorithm and write

$$s = mn + r,$$

where $0 \leqq m$ and $0 \leqq r < n$. The positive integer m has a decimal expansion as above. The tenths digit in the decimal expansion of r/n, and therefore of s/n, is b_1, where

$$10r = b_1 n + r_1, \quad 0 \leqq r_1 < n.$$

The hundredths digit is b_2, where

$$10r_1 = b_2 n + r_2, \quad 0 \leqq r_2 < n$$

and, in general, for any positive integer j, the jth digit to the right of the decimal point in the expansion of s/n is b_j, where

$$10r_{j-1} = b_j n + r_j, \quad 0 \leqq r_j < n.$$

It remains to show that only a finite number of the above steps are necessary to obtain the representation of s/n in decimal notation.

Formally, there exists a countably infinite set of remainders r_j $(j = 1, 2, \ldots)$ such that $0 \leqq r_j < n$. However, since each r_j is an element of the set $0, 1, 2, \ldots, n-1$, there are only a finite number of distinct values of the remainders r_j. Accordingly, there must exist integers p, and $q = p + t$, where t is positive, such that $r_{p-1} = r_{q-1}$, and therefore

$$\begin{aligned} nb_p + r_p &= nb_q + r_q, \\ n(b_p - b_q) &= r_q - r_p. \end{aligned}$$

Since the positive integers are well ordered, there exist least positive integers p and t having the above properties. Any $p_1 > p$ and any positive integral multiple of t also have these properties. If $b_p = b_q$, then $r_q = r_p$. If $b_p \neq b_q$, then n divides $r_q - r_p$, where $0 \leqq r_q < n$, $0 \leqq r_p < n$, and therefore $r_q - r_p = 0 = b_p - b_q$, contrary to our assumption that $b_p \neq b_q$, that is, we must have $b_p = b_q$ and $r_p = r_q$. Thus we have proved that there exist distinct positive integers p, q such that $r_{p-1} = r_{q-1}$ and that this equality implies $r_p = r_q$, that is, $r_p = r_{p+t}$, where $q = p + t$. Finally, by the principle of mathematical induction, $r_j = r_{j+t}$ for all $j \geqq p - 1$ and the positive rational number s/n may be written in the form

$$a_k 10^k + \cdots + a_1 10 + a_0 + \frac{b_1}{10} + \frac{b_2}{10^2} + \cdots + \frac{b_p}{10^{p-1}} + \frac{c_1}{10^p} + \frac{c_2}{10^{p+1}} + \cdots$$

$$+ \frac{c_t}{10^{p+t-1}} + \frac{c_1}{10^{p+t}} + \frac{c_2}{10^{p+t+1}} + \cdots + \frac{c_t}{10^{p+2t-1}} + \frac{c_1}{10^{p+2t}} + \cdots .$$

We have given a complete proof of the fact that every positive rational number may be represented by a repeating decimal. In practice, as in theory, one proceeds as above to find the r_j until some r_j is zero or is equal to some r_k, where $k < j$. For example, given $\frac{153}{7}$, we consider

$$\begin{aligned} 153 &= 21 \cdot 7 + 6, \\ 60 &= 8 \cdot 7 + 4, \\ 40 &= 5 \cdot 7 + 5, \\ 50 &= 7 \cdot 7 + 1, \\ 10 &= 1 \cdot 7 + 3, \\ 30 &= 4 \cdot 7 + 2, \\ 20 &= 2 \cdot 7 + 6, \end{aligned}$$

whence $\frac{153}{7} = 21.857142857142\ldots$. Since signed decimals are used to represent signed numbers, all rational numbers may be represented as repeating decimals. Conversely, given any decimal d in which t digits are repeated over and over, we may compute the terminating decimal $10^t d - d$ (Section 1–10) and express d as a rational number. Thus we have now shown that every rational number may be expressed as a repeating decimal, and conversely. The remaining topics considered in this chapter are important in the theory of numbers but may be omitted without disturbing the organization of this text.

EXERCISES

1. Use the above method of successive remainders and express each of the following rational numbers in decimal notation: $\frac{1}{16}, \frac{1}{50}, \frac{3}{11}, \frac{2}{3}, \frac{1}{10}, \frac{1}{128}$.
2. Repeat Exercise 1 for $\frac{17}{3}, \frac{25}{8}, \frac{11}{6}, \frac{75}{16}, \frac{125}{36}$.
3. Prove that every rational number may be represented as a repeating decimal, using the fact that at most q different remainders are obtained from the quotients $10^j/q$, where $j = 0, 1, 2, \ldots$.
4. Discuss the representation of rational numbers in the binary number system.

***2–8 Congruences.** We now proceed to divide the set of integers into subsets or subclasses with reference to an arbitrary given integer m. For example, three hours from now, fifty-one hours from now, twenty-one hours ago, and, in general, $3 + 24k$ hours from now for any integer k, all represent the same time of day. The numbers of hours $3, 51, -21, 3 + 24k$ are equivalent in a certain sense with respect to a twenty-four hour day. We say that the numbers are congruent modulo 24 and write $3 \equiv 51 \pmod{24}$. Similarly, angles of 30°, 390°, −330°, 750°, and, in general, $30° + (360k)°$ for any integer k may be represented graphically using the same initial and terminal sides. Moreover, whenever angles of $a°$ and $b°$ may be represented using the same initial and terminal sides, we have $a = b + 360k$ for some integer k and may write $a \equiv b \pmod{360}$. In general, two integers a and b are said to be *congruent modulo an integer* m if and only if there exists an integer c satisfying $a = b + cm$. Whenever such an integer c exists, we may write $a \equiv b \pmod{m}$ and call m the *modulus* of the congruence. This definition is equivalent to the statement that $a \equiv b \pmod{m}$ if and only if $a - b$ is divisible by m. For example, $3 \equiv 8 \pmod{5}$, $-3 \equiv 9 \pmod{6}$, and any two even integers are congruent modulo 2.

Congruence modulo m is an equivalence relation (Section 1–3), since it is reflexive, symmetric, and transitive, i.e.,

(i) $a \equiv a \pmod{m}$,
(ii) $a \equiv b \pmod{m}$ implies $b \equiv a \pmod{m}$, and
(iii) $a \equiv b \pmod{m}$ and $b \equiv c \pmod{m}$ imply $a \equiv c \pmod{m}$.

These three properties are easily proved on the basis of the above definition as follows: $a = a + 0m$; if $a = b + km$, then $b = a + (-k)m$; if $a = b + km$ and $b = c + hm$, then $a = c + (h + k)m$. The equivalence relation $=$ may be considered a special case of $\equiv$, where

$m = 0$. However, we shall assume $m \neq 0$ throughout our discussion.

We next consider some of the properties of this new relation, $\equiv \pmod m$. Congruences modulo m may be combined under the ring operations of addition, subtraction, and multiplication, i.e., if $a \equiv b \pmod m$ and $c \equiv d \pmod m$, then

$$a + c \equiv b + d \pmod m, \tag{2-16}$$

$$a - c \equiv b - d \pmod m, \tag{2-17}$$

$$ac \equiv bd \pmod m. \tag{2-18}$$

These congruences can be directly proved from the definition of congruence. If $a = b + sm$ and $c = d + tm$, then

$$a + c = b + d + (s + t)m,$$
$$a - c = b - d + (s - t)m,$$

and

$$ac = bd + (bt + sd + stm)m,$$

where $s + t$, $s - t$, and $bt + sd + stm$ are integers, since the set of integers is closed under the ring operations. The congruence (2-18) may be specialized in the case $a = c$, $b = d$ to give (by mathematical induction) for any positive integer n

$$a^n \equiv b^n \pmod m. \tag{2-19}$$

For example, the congruences $2 \equiv 7 \pmod 5$ and $3 \equiv 8 \pmod 5$ may be added to give $5 \equiv 15 \pmod 5$, subtracted to give $-1 \equiv -1 \pmod 5$, and multiplied to give $6 \equiv 56 \pmod 5$. Both sides of the congruence $2 \equiv 7 \pmod 5$ may be squared to give $4 \equiv 49 \pmod 5$.

Since only ring operations are required in the formation of a polynomial (Section 3-2), the congruences (2-16), (2-17), (2-18), (2-19) may now be used to give

THEOREM 2-14. *If $a \equiv b \pmod m$ and $f(x)$ is a polynomial with integral coefficients, then $f(a) \equiv f(b) \pmod m$.*

Consider as an example of this theorem the polynomial $f(x) = x^3 - 3x^2 + 2x + 1$ and the congruence $2 \equiv -1 \pmod 3$. $f(2) = 8 - 12 + 4 + 1 = 1$ and $f(-1) = -1 - 3 - 2 + 1 = -5 \equiv 1 \pmod 3$.

There is also a cancellation rule for congruences. If $ak \equiv bk \pmod m$, their difference is a multiple of m, say $(a - b)k = tm$. Let $d = (k, m)$, then

$$(a - b)k/d = t(m/d) \text{ and } a \equiv b \pmod{m/d}.$$

For example, $2 \equiv 8 \pmod 6$ implies $1 \equiv 4 \pmod 3$; $12 \equiv 32 \pmod{10}$ implies $6 \equiv 16 \pmod 5$ and $3 \equiv 8 \pmod 5$. If $ak \equiv bk \pmod m$ and $(k, m) = 1$, then $a \equiv b \pmod m$. In general, we have

THEOREM 2-15. *If $ak \equiv bk \pmod m$ and $(k, m) = d$, then $a \equiv b \pmod{m_1}$, where $m = dm_1$.*

Congruences may be used to give tests for the divisibility of any integer n by an integer m. Every positive integer can be uniquely expressed (Section 2-6) in the form

$$n = a_0 + a_1 \cdot 10 + a_2 \cdot 10^2 + \cdots + a_k \cdot 10^k, \tag{2-20}$$

where $0 \leq a_i \leq 9$ for $i = 1, 2, \ldots, k$. The familiar test for divisibility by 2 is found by considering both sides of (2-20) modulo 2. Since $10 \equiv 0 \pmod 2$, we may replace $10, 10^2, \ldots,$ and 10^k by 0 when n in (2-20) is considered modulo 2. Then $n \equiv a_0 \pmod 2$, that is, $n = a_0 + 2t$ for some integer t, and n is divisible by 2 if and only if a_0 is divisible by 2. Similarly, from $10 \equiv 1 \pmod 3$ and (2-19), we have $n \equiv a_0 + a_1 + \cdots + a_k \pmod 3$, that is, n is divisible by 3 if and only if the sum of its digits is divisible by 3. Since $10^2 \equiv 0 \pmod 4$, we have $n \equiv a_0 + 10a_1 \pmod 4$, whence n is divisible by 4 if and only if the number composed of its last two digits is divisible by 4. We may also use (2-20) to obtain

$$\begin{aligned}
&n \equiv a_0 \pmod 5,\\
&n \equiv a_0 + 3a_1 + 2a_2 - a_3 - 3a_4 - 2a_5 + a_6\\
&\qquad + 3a_7 + 2a_8 - a_9 - \cdots \pmod 7,\\
&n \equiv a_0 + a_1 + \cdots + a_k \pmod 9,\\
&n \equiv a_0 - a_1 + a_2 - a_3 + \cdots + (-1)^k a_k \pmod{11},\\
&n \equiv a_0 - 3a_1 - 4a_2 - a_3 + 3a_4 + 4a_5 + a_6\\
&\qquad - 3a_7 - 4a_8 - a_9 + \cdots \pmod{13},\\
&n \equiv a_0 + 10a_1 \pmod{25},
\end{aligned} \tag{2-21}$$

and many other such tests for divisibility. For example, $342538 \equiv 0 \pmod 7$, i.e., it is divisible by 7, since $a_0 = 8$, $a_1 = 3$, $a_2 = 5$, $a_3 = 2$, $a_4 = 4$, $a_5 = 3$, and, using the test given above, $342538 \equiv 8 + 3 \cdot 3 + 2 \cdot 5 - 2 - 3 \cdot 4 - 2 \cdot 3 = 7 \equiv 0 \pmod 7$. Similarly, 3637425 is divisible by 11, and 7587125 is divisible by 13. The periodic nature of the multiples of the digits a_i is considered in Exercises 5 and 6, Section 2-10.

Before the invention of the calculating machine, many arithmetical processes were checked by the method of *casting out nines*, i.e., the

numbers were considered modulo 9 as in (2–21) and the congruences (2–16), (2–17), (2–18) were used. The product $321 \cdot 152 = 48792$ would be checked by the congruences $321 \equiv 6 \pmod 9$, $152 \equiv 8 \pmod 9$, $321 \cdot 152 \equiv 6 \cdot 8 \equiv 48 \equiv 3 \equiv 48792 \pmod 9$. This method is not a complete check, since some errors, such as the interchange of two digits, are not located. In the case of a quotient $a/b = q + r/b$, the relation $a = qb + r$ must hold and one checks that $a \equiv qb + r \pmod 9$. For example, the equation $\frac{83}{17} = 4 + \frac{15}{17}$ is checked by considering $83 = 4 \cdot 17 + 15$, which becomes $2 \equiv 4 \cdot (-1) + 6 \pmod 9$.

Exercises

1. Prove that $a^2 \equiv 1 \pmod 8$, where a is any odd number.
2. Give four examples for Theorem 2–14, using polynomials of degree at least three.
3. Find $f(13) \pmod 9$ where $f(x) = 7x^5 + 13x^4 - 72x^3 + 2153$.
4. Give three numerical examples illustrating Theorem 2–15.
5. Give tests for divisibility by 6, 8, and 15.
6. Use congruence theorems to test 1113 and 23,535 for divisibility by 7, 9, 11, and 15.
7. Test the following by the method of casting out nines:
   (a) $1235 \cdot 341 = 421135$,
   (b) $852 + 1239 + 251 + 172 = 2514$.
8. Develop a method of casting out elevens. Repeat Exercise 7, casting out elevens.
9. Find a test for divisibility by 4 when numbers are expressed to the base five. Give one three-digit and one four-digit example.
10. Find a test for divisibility by $n - 1$ when numbers are expressed to the base n.
11. Find a test for divisibility by $n + 1$ when numbers are expressed to the base n.
12. Develop tests for divisibility by 4, 8, and 16, and prove that at most b digits need to be considered to test any given integer expressed to the base 10 for divisibility by 2^b.
13. Prove that $a \equiv b \pmod m$, $0 < a < m$, $0 < b < m$ imply $a = b$.
14. Three brothers decided at school to divide their common box of marbles among the seven members of their gang. The first boy home divided the marbles into seven piles and had one marble left over. He took his pile and the extra marble. The second boy home divided the remaining marbles into seven piles, had one marble left over, gave the extra marble to his sister, and took his pile. When the third boy came home, he divided the remaining marbles evenly into seven piles. Find the smallest possible number of marbles that could have been in the box originally. Give another possible answer for the original number of marbles. Give all possible solutions as a congruence class (Section 2–9). (*Hint:* Start by expressing the desired number to the base seven, say, $N = abc_7$.)

***2–9 Residue Classes. Euler's ϕ-function.** Given any two integers c and m, the Division Algorithm states that there exist integers q and r, $0 \leq r < m$, such that $c = qm + r$. The number r is called the *residue* of c modulo m and we write $c \equiv r \pmod m$. For example, $7 \equiv 2 \pmod 5$, $31 \equiv 1 \pmod 5$, $102 \equiv 2 \pmod 5$. The totality of numbers c such that $c \equiv r \pmod m$ is called a *residue class* or *congruence class* modulo m and is indicated by $[r] \pmod m$. The numbers in the class $[r] \pmod m$ are

$$r, r \pm m, r \pm 2m, r \pm 3m, \ldots .$$

For example, the residue class $[2] \pmod 5$ consists of the numbers

$$\ldots, -13, -8, -3, 2, 7, 12, 17, \ldots .$$

Every integer, positive, negative, or zero, belongs modulo 5 to one of the residue classes [0], [1], [2], [3], [4] (mod 5). In general, every integer belongs modulo m to one of the residue classes [0], [1], [2], $\ldots$, $[m - 1] \pmod m$ (see Exercise 8). For $m = 2$ all even numbers are in the residue class [0] (mod 2) and all odd numbers are in the residue class [1] (mod 2). A set of numbers $r_1, r_2, \ldots, r_m$, one of which lies in each of the classes [0], [1], [2], $\ldots$, $[m - 1] \pmod m$, is called a *complete residue system* modulo m. For example, the numbers 5 and 8 form a complete residue system modulo 2; the numbers 64, 17, 34, and -1 form a complete residue system modulo 4.

A residue class $[r] \pmod m$ may be expressed in terms of any one of its elements, that is, $[r] \pmod m = [r + km] \pmod m$ for any integer k. Thus for any integer m the total number of distinct residue classes modulo m is $|m|$. A set of numbers $r_1, r_2, \ldots, r_m$ is a complete residue system modulo m if and only if $r_i \not\equiv r_j \pmod m$ whenever $i \neq j$ and $i, j = 1, 2, \ldots, m$.

Complete residue systems may be used in the determination of all nth roots of unity from a given primitive nth root of unity. By definition (Section 1–17), s is a primitive nth root of unity if and only if n is the smallest positive integer k such that $s^k = 1$. Then if $s^m = 1$, we may write $m = qn + r$, where $0 \leq r < n$, and obtain $s^m = s^{qn+r} = s^{qn} \cdot s^r = s^r = 1$. Now since $0 \leq r < n$, $s^r = 1$ and n is the smallest positive integer k such that $s^k = 1$, we have $r = 0$ and $m = qn$, that

is, $s^m = 1$ if and only if $m \equiv 0 \pmod{n}$, where s is a primitive nth root of unity.

The existence of at least one primitive nth root of unity for any positive integer n is assured by De Moivre's Theorem (Section 1–17). Given a primitive nth root of unity s, we found in Section 1–17 that every integral power of s, say s^t, was also an nth root, since $(s^t)^n = (s^n)^t = 1^t = 1$. Furthermore, if $s^t = s^u$, then $s^{t-u} = 1$ and $t - u \equiv 0 \pmod{n}$, that is, $t \equiv u \pmod{n}$. Thus two integral powers of a primitive nth root of unity are distinct if and only if the exponents are from distinct residue classes modulo n. Using these facts, we may now generalize Theorem 1–7 as follows:

THEOREM 2–16. *All nth roots of unity are given by the sequence*

$$s^{r_1}, s^{r_2}, \ldots, s^{r_n},$$

where s is a primitive nth root of unity and the r_i's form a complete residue system modulo n.

For example, if $n = 4$, then 16, -11, 30, 67 form a complete residue system and all the fourth roots of unity are in the set $i^{16} = 1$, $i^{-11} = i$, $i^{30} = -1$, $i^{67} = -i$, where $i = \sqrt{-1}$.

We now prepare to define a second type of residue system, called a reduced residue system.

For any positive integer m the number of positive integers less than or equal to m and relatively prime to m is denoted by $\phi(m)$ and is called the indicator (totient) of m or *Euler's ϕ-function* of m. Thus $\phi(m)$ is the number of integers k in the set

$$1, 2, 3, \ldots, m - 1, m \tag{2–22}$$

such that $(k, m) = 1$; $\phi(2) = 1$, $\phi(3) = 2$, $\phi(4) = 2$, $\phi(5) = 4$, $\phi(6) = 2$, ... From the above and the definition of relatively prime (Section 2–1) we have $\phi(1) = 1$.

If $(r, m) = 1$, then for any integer k we have $(r + km, m) = 1$, that is, every element of $[r] \pmod{m}$ is relatively prime to m. Thus a residue class $[r] \pmod{m}$ is *relatively prime* to m if and only if $(r, m) = 1$. We use this fact in our definition of a reduced residue system. A set of numbers $r_1, r_2, \ldots, r_{\phi(m)}$, one of which lies in each residue class that is relatively prime to m, is called a *reduced residue system* modulo m.

This second type of residue system may also be used in the study of nth roots of unity. Suppose s is a primitive nth root of unity,

and consider the conditions upon k such that s^k will be a primitive nth root of unity. We have $(s^k)^n = 1$ for any integer k, since s is a primitive nth root. If $(k, n) = d > 1$, then $k = k_1 d$, $n = n_1 d$, where $n_1 < n$ and $(s^k)^{n_1} = (s^{k_1 d})^{n_1} = (s^{n_1 d})^{k_1} = 1^{k_1} = 1$, that is, s^k is not a primitive nth root if $(k, n) = d > 1$. If $(k, n) = 1$ and $(s^k)^m = 1$, then $km \equiv 0 \pmod{n}$; that is, $n | km$ and, by Theorem 2–9, $n | m$, whence s^k is a primitive nth root of unity. Thus s^k is a primitive nth root of unity if and only if $(k, n) = 1$. This fact enables us to find all primitive nth roots from a given primitive nth root of unity.

THEOREM 2–17. *If s is a primitive nth root of unity, then all primitive nth roots of unity are in the set*

$$s^{r_1}, s^{r_2}, \ldots, s^{r_{\phi(n)}},$$

where the r's form a reduced residue system modulo n.

We shall also use reduced residue systems in the proofs of Euler's Theorem and Fermat's Simple Theorem (Section 2–10).

EXERCISES

1. Write down complete residue systems modulo each of the following integers: 4, 5, 9, 11, and 16.

2. Write down reduced residue systems modulo each of the following integers: 4, 5, 9, 11, 16, 31, 60, -70.

3. Prove that the numbers -5, -2, 12, 26, 39, 53 form a complete residue system modulo 6.

4. Prove that any m consecutive integers form a complete residue system modulo m.

5. Prove that if $(d, m) = 1$, then $d, 2d, 3d, \ldots, md$ form a complete residue system modulo m.

6. Prove that $a + r_1$, $a + r_2$, $a + r_3, \ldots, a + r_m$ is a complete residue system modulo m for any integer a if $r_1, r_2, r_3, \ldots, r_m$ is a complete residue system modulo m.

7. Express a complete residue system modulo mn where $(m, n) = 1$ in terms of given complete residue systems of m and n.

8. Use the Division Algorithm and prove that every integer belongs modulo m to one and only one of the residue classes $[0], [1], \ldots, [m - 1] \pmod{m}$ for an arbitrary integer $m \neq 0$.

9. Define $[a] + [b] = [a + b]$ modulo m, $[a] \cdot [b] = [ab]$ modulo m, and prove that these definitions are independent of the particular elements a, b selected from the residue classes $[a], [b]$ modulo m.

10. Prove that the residue classes modulo 5 form a ring.

11. Prove that the residue classes modulo 6 form a ring.

12. Prove that the residue classes modulo m for any integer $m \neq 0$ form a ring.

13. Use residue classes modulo 6 to illustrate the concept of zero divisors (Section 1–14).

14. Prove that the residue classes modulo 5 form a field.

15. Prove that the residue classes modulo p for any prime number p form a field.

***2–10 Evaluation of $\phi(m)$.** Since every positive integer can be uniquely expressed as a product of prime numbers (Theorem 2–8), we first evaluate the ϕ-function for prime numbers. If m is a prime number, then every number except m in (2–22) is relatively prime to m and $\phi(m) = m - 1$. If $m = p^a$, where p is a prime, then in the set $1, 2, 3, \ldots, p, p+1, \ldots, 2p, 2p+1, \ldots, 3p, \ldots, p^a$, only the numbers $p, 2p, 3p, \ldots, (p^{a-1})p$ are divisible by p. Thus $p^a - p^{a-1}$ of the numbers are prime to p (Theorem 2–5) and

$$\phi(p^a) = p^a\left(1 - \frac{1}{p}\right).$$

We next prove that if $m = uv$, where $(u, v) = 1$, then $\phi(m) = \phi(u)\phi(v)$ and, in general, the ϕ-function of a product of relatively prime factors is equal to the product of the ϕ-functions of the factors. This can be quickly proved by using the fact that there are exactly $\phi(m)$ primitive mth roots of unity and that all the mth roots of unity form a cyclic group (Section 1–17). A longer but more elementary proof is obtained by writing down all integers $1, 2, 3, \ldots, uv$ in a rectangular array as follows:

1	2	3	$\ldots$	h	$\ldots$	u
$u+1$	$u+2$	$u+3$	$\ldots$	$u+h$	$\ldots$	$2u$
$2u+1$	$2u+2$	$2u+3$	$\ldots$	$2u+h$	$\ldots$	$3u$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$	$\cdot$
$(v-1)u+1$	$(v-1)u+2$	$(v-1)u+3$	$\ldots$	$(v-1)u+h$	$\ldots$	$vu.$

For example, if $u = 5$ and $v = 3$, we write

1	2	3	4	5
6	7	8	9	10
11	12	13	14	15

Each row forms a complete residue system modulo u. Each column forms part of a single residue class (mod u), i.e., every element of the column headed by a number h belongs to $[h]$ (mod u). Thus the elements in the column headed by h are relatively prime to u if and only if $(h, u) = 1$. The number of columns whose elements are relatively prime to u is therefore $\phi(u)$. Next we prove that in each column no two of the elements belong to the same residue class modulo v. Consider $su + h$ and $tu + h$. By the Division Algorithm,

$$su + h = q_s v + r_s, \quad 0 \leqq r_s < v,$$
$$tu + h = q_t v + r_t, \quad 0 \leqq r_t < v.$$

Suppose $r_s = r_t$, then $(s - t)u = (q_s - q_t)v$. Since $(u, v) = 1$ and $u > 0$, either $q_s = q_t$ or v divides $s - t$. But $s < v$, $t < v$ and, using Exercise 3, Section 2–1, we have $s = t$, that is, no two distinct elements of any fixed column are congruent modulo v. Since there are v elements in each column, each column constitutes a complete residue system modulo v and contains exactly $\phi(v)$ elements that are relatively prime to v. Therefore, on each of the $\phi(u)$ columns of elements relatively prime to u there are $\phi(v)$ elements that are also relatively prime to v, that is, there are $\phi(u)\phi(v)$ elements relatively prime to both u and v and, therefore, to uv (Theorem 2–9). In other words, $\phi(uv) = \phi(u)\phi(v)$. In general, if $m_1, m_2, \ldots, m_k$ are k positive integers which are relatively prime each to each, then

$$\phi(m_1 m_2 \ldots m_k) = \phi(m_1)\phi(m_2) \ldots \phi(m_k).$$

The last two paragraphs and Theorem 2–8 now give us

THEOREM 2–18. *For any positive integer $m = p_1^{a_1} \cdot p_2^{a_2} \ldots p_n^{a_n}$, where the p's are distinct prime numbers,*

$$\phi(m) = m\left(1 - \frac{1}{p_1}\right)\left(1 - \frac{1}{p_2}\right) \cdots \left(1 - \frac{1}{p_n}\right).$$

For example: $\phi(15) = 15(1 - \frac{1}{5})(1 - \frac{1}{3}) = 8$

or $\phi(15) = \phi(3)\phi(5) = 2 \cdot 4 = 8.$

We also have

THEOREM 2–19. *Given positive integers m, n, d such that $m = nd$, the number of integers $k \leqq m$ such that $(k, m) = d$ is $\phi(n)$.*

This is easily proved, since every number $\leqq m = nd$ having a divisor d is one of the set

$$d, 2d, 3d, \ldots, td, \ldots, (n-1)d, nd$$

and $(td, m) = d$ or $(td, nd) = d$ if and only if $(t, n) = 1$. Thus the number of values of t such that $(td, m) = d$ is exactly $\phi(n)$.

The value of $\phi(m)$ for any positive integer m can be found using Theorem 2–18. For $m \leqq 10{,}000$ these values are available in a set of tables by J. W. L. Glaisher.

A very important application [43; 272–310] of the ϕ-function is given by the following theorem.

THEOREM 2–20. EULER'S THEOREM. *If m is a positive integer and a is any integer such that $(a, m) = 1$, then $a^{\phi(m)} \equiv 1 \pmod{m}$.*

If m is a prime number p, this theorem reduces to a theorem stated earlier by Fermat.

THEOREM 2–21. FERMAT'S SIMPLE THEOREM. *If p is a prime number and $p \nmid a$, then $a^{p-1} \equiv 1 \pmod{p}$.*

Theorem 2–21 is frequently expressed in the form $a^p \equiv a \pmod{p}$ which holds for all positive integers a.

One proof of Theorem 2–20 and therefore also of Theorem 2–21 involves a reduced residue system modulo m, say $r_1, r_2, \ldots, r_{\phi(m)}$. Since by hypothesis $(a, m) = 1$, the set of elements $ar_1, ar_2, \ldots, ar_{\phi(m)}$ also constitutes a reduced residue system modulo m. The elements of the two systems must therefore be congruent modulo m in pairs (under some ordering), and we have by repeated application of (2–18)

$$a^{\phi(m)} r_1 r_2 \ldots r_{\phi(m)} \equiv r_1 r_2 \ldots r_{\phi(m)} \pmod{m}.$$

By definition of a reduced residue system, $(r_i, m) = 1$ $[i = 1, 2, \ldots, \phi(m)]$. Thus by Theorem 2–15, we may divide both sides of the above equation by $r_1 r_2 \ldots r_{\phi(m)}$ and obtain $a^{\phi(m)} \equiv 1 \pmod{m}$. This completes our proof of Theorem 2–20 and also of Theorem 2–21. In the remaining two sections of this chapter we shall consider linear congruences and Diophantine problems.

EXERCISES

1. Find $\phi(12)$, $\phi(32)$, $\phi(17)$, $\phi(31)$, $\phi(60)$.
2. Prove that $(n-1)! \equiv 0 \pmod{n}$, where n is any composite number different from 4.
3. Prove that $(a+b)^p \equiv a^p + b^p \pmod{p}$, where a and b are any integers and p is any prime number.
4. Verify Euler's Theorem when $a = 7$ and $m = 12$.

5. If $(m, 10) = 1$, prove that in the test for divisibility by m of any sufficiently large number expressed to the base 10, the multiples of the digits must occur in sets very much like the digits of a repeating decimal. For example, $10^2 \equiv 1 \pmod{11}$ and $n \equiv a_0 - a_1 + a_2 - a_3 + \cdots \pmod{11}$ where the multiples are the set $+1, -1$ repeated until all digits of the given number have been considered.
6. If $(m, 10) \neq 1$, write $m = 2^a 5^b n$ and use Exercise 3, Section 2–7, to show that for any sufficiently large number m the multiples of the digits occur in repeating sets after a certain finite number of digits have been considered.

***2–11 Linear congruences.** Arithmetic is primarily concerned with numbers. In algebra, new symbols called *variables* (Section 3–1) are introduced. We now turn momentarily from arithmetic to algebra. As in the theory of equations, we may consider the problem of finding integers x that satisfy a polynomial congruence $f(x) \equiv 0 \pmod{m}$. If a is an integer such that $f(a) \equiv 0 \pmod{m}$ and $a \equiv b \pmod{m}$, then by Theorem 2–14, we have $f(b) \equiv 0 \pmod{m}$ and there is a countably infinite set of integers

$$a, a \pm m, a \pm 2m, a \pm 3m, \ldots$$

satisfying the congruence $f(x) \equiv 0 \pmod{m}$. One speaks of the whole residue class $[a] \pmod{m}$ as a single *solution* of the polynomial congruence. Thus the number of solutions of $f(x) \equiv 0 \pmod{m}$ is the number of residue classes $[r_1], [r_2], \ldots, [r_k] \pmod{m}$ such that $f(r_i) \equiv 0 \pmod{m}$. Since there are exactly m distinct residue classes, any given polynomial congruence has at most m solutions modulo m.

In the ring of integers, both sides of an equation $ax = b$ may be divided (Section 2–1) by a if and only if there exists an integer c such that $ac = b$. Similarly, the divisibility of both sides of a congruence modulo m by an integer is related to the solution of a linear congruence $ax \equiv b \pmod{m}$. Thus we seek integral values of the variable x that satisfy

$$ax \equiv b \pmod{m}. \tag{2–23}$$

Let us first consider a special case of (2–23). The congruence

$$ax \equiv 1 \pmod{m} \tag{2–24}$$

holds if and only if $ax = 1 + km$ or $ax - km = 1$ for some integer k. Then, by Theorem 2–12, $ax \equiv 1 \pmod{m}$ if and only if $(a, m) = 1$. Thus (2–24) has a unique solution if and only if $(a, m) = 1$. When

(2–24) has a unique solution $[b]$ (mod m) any element of $[b]$ is called a *reciprocal* a^{-1} of a modulo m. Accordingly, an integer a has a reciprocal modulo m if and only if $(a, m) = 1$. For example: 3 is a reciprocal of 2 modulo 5, 4 is a reciprocal of 2 modulo 7, but 2 has no reciprocal modulo 6. If such exists, a reciprocal of n modulo m can always be found by Theorem 2–12, since if $(n, m) = 1$ there exist integers A and B such that $Am + Bn = 1$, and B is the reciprocal of n modulo m.

We now consider (2–23). If $(a, m) = 1$, then a has a reciprocal a^{-1}, and we may write $a^{-1}ax \equiv a^{-1}b \pmod{m}$, $x \equiv a^{-1}b \pmod{m}$. Thus (2–23) has a solution if $(a, m) = 1$. In general, if $(a, m)|b$, we let $(a, m) = d$, $a = a_1d$, $m = m_1d$, and $b = b_1d$. Then $(a_1, m_1) = 1$ and $a_1x \equiv b_1 \pmod{m_1}$ has a solution $x \equiv a_1^{-1}b_1 \pmod{m_1}$, that is, $a_1x = b_1 + km_1$ for some integer k. From this equation, we get $a_1dx = b_1d + km_1d$, or $ax = b + km$, whence (2–23) has a solution if $(a, m)|b$. Conversely, if $ax \equiv b \pmod{m}$ has a solution $[r] \pmod{m}$, then $ar = b + km$ for some integer k, whence $ar - km = b$ and $(a, m)|b$. Thus $ax \equiv b \pmod{m}$ is always solvable if $(a, m) = 1$ and, in general, we have

THEOREM 2–22. *The congruence* $ax \equiv b \pmod{m}$ *is solvable if and only if* $d = (a, m)$ *divides* b.

For example, $2x \equiv 1 \pmod{6}$ and $3x \equiv 5 \pmod{6}$ have no solutions; $2x \equiv 1 \pmod{5}$ and $3x \equiv 9 \pmod{6}$ are solvable.

If $ax \equiv b \pmod{m}$ and $ay \equiv b \pmod{m}$, then $ax \equiv ay \pmod{m}$, $ax = ay + km$. As before, let $d = (a, m)$, $a = da_1$, $m = dm_1$. Then $a_1(x - y) = km_1$ and $x \equiv y \pmod{m_1}$. Thus any two solutions of (2–23) are congruent modulo m_1. If $[x] \pmod{m}$ is a solution of (2–23), then using $b = db_1$, we have $a_1dx = b_1d + kdm_1$, whence $a_1x = b_1 + km_1$, that is, any solution of (2–23) is a solution of $a_1x \equiv b_1 \pmod{m_1}$. By Theorem 2–22, $a_1x \equiv b_1 \pmod{m_1}$ has a solution $[x_0] \pmod{m_1}$ that by the preceding argument is unique. Thus all solutions of (2–23) are in $[x_0] \pmod{m_1}$, that is, in the set

$$x_0, x_0 \pm m_1, x_0 \pm 2m_1, \ldots, x_0 \pm dm_1, \ldots.$$

Since $x_0 + dm_1 \equiv x_0 \pmod{m}$, there are exactly d solutions (mod m), namely, $[x_0], [x_0 + m_1], \ldots, [x_0 + (d - 1)m_1] \pmod{m}$. Thus we have

THEOREM 2–23. *If* $ax \equiv b \pmod{m}$ *has a solution, then there is one unique solution* (mod m/d) *where* $(a, m) = d$, *and* d *solutions* (mod m).

For example, $6x \equiv 9 \pmod{15}$ has a unique solution $[4] \pmod{5}$ and three solutions [4], [9], [14] (mod 15), where $d = (6, 15) = 3$.

Results similar to the above may be obtained for simultaneous congruences involving several moduli [43; 240–249]. In particular, the *Chinese Remainder Theorem* [If integers $m_1, m_2, \ldots, m_t$ are relatively prime in pairs, there exist integers x for which simultaneously $x \equiv a_1 \pmod{m_1}$, $x \equiv a_2 \pmod{m_2}$, $\ldots$, $x \equiv a_t \pmod{m_t}$] is the basis for many interesting problems. However, in a brief treatment many details must be omitted. Thus we consider our aim of introducing basic concepts of linear congruences accomplished and leave such interesting applications as the one mentioned above for the reader to pursue in texts devoted entirely to the theory of numbers.

EXERCISES

1. Prove that factorization is not necessarily unique (mod m) where m is not a prime number by exhibiting two distinct factorizations of $x^2 - 1$ modulo 15.
2. Solve the congruences
   (a) $x^2 - 6x + 5 \equiv 0 \pmod{4}$,
   (b) $x^3 + 2x^2 + 4x + 3 \equiv 0 \pmod{5}$.
3. How many solutions has the congruence $x^{17} \equiv x \pmod{17}$?
4. Solve the congruence $x^7 + 2x^6 + 8x^5 + x + 3 \equiv 0 \pmod{5}$.
5. Find the reciprocals of 7 (mod 13), 5 (mod 33), 12 (mod 49).
6. Solve if possible:
   (a) $4x \equiv 1 \pmod{5}$,
   (b) $4x \equiv 1 \pmod{6}$,
   (c) $6x \equiv 39 \pmod{15}$,
   (d) $6x \equiv 39 \pmod{34}$,
   (e) $1250x \equiv 1725 \pmod{2000}$.
7. Find all solutions of the following congruences:
   (a) $4x \equiv 6 \pmod{10}$,
   (b) $10x \equiv 8 \pmod{16}$.
8. Prove Wilson's Theorem: $(p - 1)! \equiv -1 \pmod{p}$, for any prime number p.
9. Prove [43; 241–242] that two congruences $x \equiv a \pmod{m}$ and $x \equiv b \pmod{n}$ have a common solution if and only if $a \equiv b \pmod{s}$, where $s = (m, n)$. Give a method for finding the solution when such exists.

***2–12 Diophantine problems.** We conclude our brief study of the theory of numbers by mentioning two famous problems. The first concerns the integral solutions of the *Pythagorean equation*, $a^2 + b^2 = c^2$; the second is known as Fermat's Last Theorem. Both problems are concerned with integral solutions and may be called

Diophantine problems, i.e., algebraic problems in which rational solutions are desired. Such problems are discussed in most texts on the theory of numbers, for example, [43; 165–208] and [50; 37–67, 388–428].

The Pythagorean equation is an algebraic expression of the *Pythagorean theorem*, which states essentially that in a right triangle the sum of the squares of the lengths of the legs is equal to the square of the length of the hypotenuse. The problem of finding all integral solutions of the Pythagorean equation is thus precisely the problem of finding all right triangles whose sides have integral lengths. The particular solution $a = 3, b = 4, c = 5$, along with $a = 5, b = 12, c = 13$ and $a = 8, b = 15, c = 17$, was known by early Chinese, Hindu, and Egyptian writers. A somewhat more general solution

$$a = 2n + 1, \quad b = 2n^2 + 2n, \quad c = 2n^2 + 2n + 1, \tag{2–25}$$

where n is any integer, is attributed by the Greeks to Pythagoras himself.

We may easily verify by substitution that (2–25) is a solution for any integer n. However, the relation $b + 1 = c$ which must hold for all solutions obtained from (2–25) does not need to hold for all solutions of the Pythagorean equation. For example, $a = 8$, $b = 15$, $c = 17$ is a solution which cannot be obtained from (2–25). Many other such solutions arise from the fact that if a, b, c is a solution of the Pythagorean equation, then da, db, dc is also a solution for any integer d. Thus (2–25) does not give all solutions of the Pythagorean equation or even all solutions such that a, b, and c are relatively prime, i.e., *primitive solutions*. All primitive solutions of the Pythagorean equation are given [50; 40] by the formulas

$$a = r^2 - s^2, \quad b = 2rs, \quad c = r^2 + s^2, \tag{2–26}$$

where $(r, s) = 1$, $0 < s < r$, and $r \not\equiv s \pmod 2$.

The other problem that we shall mention has been a constant challenge to mathematicians for over three hundred years.

THEOREM 2–24. FERMAT'S LAST THEOREM. *If n is an integer greater than 2, there do not exist integers x, y, z where $xyz \neq 0$ such that $x^n + y^n = z^n$.*

Fermat [43; 203–207] saw this theorem as an extension of the Pythagorean equation and indicated that he had a "truly wonderful proof" of it, but never stated the proof. Even though large prizes have been offered for a proof and the theorem has been proved for all $n \leqq 617$, no general proof has yet been found. The status of the theorem was summarized in 1946 by H. S. Vandiver [51].

Throughout this chapter we have considered properties of the ring of integers. Divisibility and the Division Algorithm were used in the discussion of prime numbers, unique factorization, and the Euclidean Algorithm. Representation of numbers to several bases has been considered. In the decimal notation it was found that every rational number may be represented as a repeating decimal and conversely. The concept of a congruence modulo an integer m was used to verify several common methods of checking divisibility and arithmetical computations. This concept was also used to subdivide the set of integers into congruence or residue classes. The concept of residue class led to that of a complete residue system modulo m consisting of exactly one element from each class. Then it was found that in any complete residue system modulo m the elements relatively prime to m constituted a reduced residue system modulo m. Given a primitive mth root of unity, all mth roots were obtained using any complete residue system modulo m, all primitive mth roots using any reduced residue system modulo m. The properties of a reduced residue system were used to prove two classical theorems in the theory of numbers, Euler's Theorem and Fermat's Simple Theorem. Finally, linear congruences and Diophantine problems were briefly mentioned. Our task has been primarily to present a few fundamental concepts and thereby in particular to give a better understanding of the properties and behavior of the integers to readers who will not have an opportunity to undertake a full course on this one phase of mathematics. In the next chapter we shall reconsider many of the properties of the ring of integers as properties of a ring of polynomials.

EXERCISES

1. Prove [50; 38–40] that all primitive solutions of the Pythagorean equation are given by (2–26).

2. List the twenty possible right triangles having all three sides of integral lengths and the longest side not over fifty units long.

3. A class of 12 children have b apples each for lunch. Another class of 8 children have c oranges each. Find two possible pairs of values of b and c such that the two classes of children may exchange lunches and distribute them equally. What are the least possible positive values of b and c? Give a complete solution of the problem using congruence classes.

4. Discuss the work and methods of Diophantos of Alexandria.

CHAPTER 3

THEORY OF POLYNOMIALS

The positive integers have been used in Chapter 1 to define rational, algebraic, transcendental, and real numbers. Properties of the ring of integers have been discussed in Chapter 2. In this chapter we shall use a ring of polynomials in one variable to define rational, algebraic, transcendental, and analytic functions. Divisibility, the Division Algorithm, the Euclidean Algorithm and properties in the ring of polynomials corresponding to prime numbers, bases, and congruences in the ring of numbers will be discussed. Our purpose is threefold: to understand the basic properties of polynomials, to see the relationships between polynomials and other common functions, and to introduce a few concepts that will be needed in our discussion of the theory of equations in Chapter 4.

3–1 Polynomials. In the first two chapters, we have been primarily concerned with numbers: integers, rational numbers, real numbers, complex numbers. We now introduce a new set of symbols $x, y, t, \ldots$ and consider equality, addition, subtraction, multiplication, and division in the total set composed of the new symbols and the complex numbers. The new symbols may be considered simply as symbols without any sets of values or assumed relations. In this case they are called *indeterminates*. The new symbols may also be considered as *variables* that take on values from a subset of the set of complex numbers. We shall usually call the symbols variables, although we shall at times mention corresponding properties of indeterminates. A great deal of the theory considered in this chapter will apply to both variables and indeterminates.

Given any indeterminate x, we define the symbol x^n for any positive integer n to represent the product of n factors x, $x^0 = 1$, $x^{-n}x^n = 1$, and $(x^{1/n})^n = x$. Similar definitions hold for any variable x, with the exception that x^0 and x^{-n} are undefined when $x = 0$. Addition and multiplication of the new symbols and complex numbers may be defined such that they are unique, commutative, associative, and satisfy the distributive laws. Thus $ax + bx = (a + b)x$ and $(ax)(bx) = abx^2$ for any complex numbers a, b.

The product of any set of complex numbers and the new symbols

is called a *monomial*. For example, 15, x, $2x$, $5x^2y^3t$, and $3\sqrt{2}xy$ are monomials. The sum of two monomials is called a *binomial*. A sum of three monomials is called a *trinomial* and, in general, a sum of one or more monomials is called a *polynomial*.

A monomial of the form bx^m, where m is a nonnegative integer and b is a complex number, is called a monomial in x with *coefficient* b and, when $b \neq 0$, of *degree* m. Any complex number b is itself a monomial. The monomial 0 has no degree. When $b \neq 0$, the monomial $b = bx^0$ has degree zero.

A polynomial of the form

$$a_0x^m + a_1x^{m-1} + \cdots + a_{m-1}x + a_m, \tag{3–1}$$

where the a_i are complex numbers and $a_0 \neq 0$, is called a polynomial of degree m in x. The a_i are called the coefficients of the polynomial. The nonzero leading coefficient a_0 is called the *initial* of the polynomial. Since indeterminates do *not* take on numbers as values, two polynomials in an indeterminate x are equal if and only if the coefficients of corresponding powers of x are equal. Thus, for an indeterminate x, the equation

$$ax^2 + bx + c = x + 2$$

implies that $a = 0$, $b = 1$, and $c = 2$. Two polynomials in a variable x may be equal for any nonnegative integral number of values of the variable x. Thus for a variable x, the equation $x^2 - 2x - 3 = 0$ implies $x = 3$ or $x = -1$.

The degree of a polynomial depends upon the variable under consideration. For example, $3x^2y^5$ is of degree two in x with coefficient $3y^5$ and of degree five in y with coefficient $3x^2$. The polynomial $10x^6$ may be considered as a polynomial of degree six in x with coefficient 10, a polynomial of degree two in $2x^3$ with coefficient $\frac{5}{2}$, a polynomial of degree twelve in $\sqrt{x}$ with coefficient 10, and in many other ways. We shall use the notation $p(x)$ to indicate a polynomial in x, $p(\sqrt{2x})$ to indicate a polynomial in $\sqrt{2x}$.

Exercises

1. List five polynomials in x.
2. Give the initial and degree of each polynomial listed in Exercise 1.
3. Find the coefficient when $36x^2$ is considered a polynomial in:
   (a) x, (d) $3x^2$,
   (b) $2x$, (e) $\sqrt{x}$, that is, y where $y^2 = x$,
   (c) x^2, (f) $\sqrt[3]{2x}$, that is, y where $y^3 = 2x$.

4. Give the degree of the polynomial in each part of Exercise 3.

5. Write $8x^3 + 12x^2 - 10x + 7$ as a polynomial in (a) $2x$, (b) $\sqrt{x}$.

6. Given two polynomials $p(x)$ and $q(x)$ of degrees m and n respectively and with complex numbers as coefficients, prove that the product of the two polynomials has degree $m + n$.

7. Repeat Exercise 6 for the product of an arbitrary finite number of polynomials with degrees m_i.

3–2 Rings of polynomials. Each polynomial (3–1) consists of a variable x and a set of coefficients $a_0, a_1, \ldots, a_m$ combined under the ring operations of addition, subtraction, and multiplication. The polynomials in a single variable x are often classified according to their coefficients. Thus we shall speak of polynomials in x with integral coefficients, polynomials in x with rational coefficients, with real coefficients, with complex coefficients. These sets of polynomials are sometimes respectively called integral polynomials, rational polynomials, real polynomials, and complex polynomials in x. Each set has the preceding sets as subsets. The sum of two polynomials with integral coefficients is a polynomial with integral coefficients. Similar statements may be made for the product and difference. Thus the polynomials in x with integral coefficients form a ring. In general, given a ring T of numbers, the set of polynomials in x with coefficients from the set of numbers T forms a *ring of polynomials*. This ring of polynomials is an integral domain [34; 33–34] if and only if T is an integral domain as defined in the introductory paragraph of Chapter 2.

When the coefficients of a set of polynomials may be arbitrary elements of a field or number system (Section 1–14), we shall be able (Section 3–5) to apply the Division Algorithm (Section 2–2) to the ring of polynomials. Accordingly, we shall be primarily concerned throughout this chapter with rings of polynomials in which the coefficients may be arbitrary elements of a field such as the rational number system, the real number system, or the complex number system.

Polynomials in several variables may be defined as arising from a finite set of variables and a set of coefficients combined under a finite set of ring operations. Although we shall usually find it convenient to discuss polynomials in a single variable, many statements will apply equally to polynomials in several variables. The most important exception is the Division Algorithm (Section 3–5) and its many applications. Except when otherwise specified, we shall hereafter consider polynomials in a single variable with arbitrary real numbers as coefficients.

Exercises

1. Describe five different rings of polynomials.

2. Describe five different integral domains of polynomials (See Exercise 9, Section 2–1).

3. Describe two rings of polynomials that are not integral domains.

3–3 Rational functions. The set of positive integers in Chapter 1 was gradually extended to the set of rational numbers, the real numbers, and the complex numbers. In this chapter we shall consider extensions of the set of polynomials to the set of rational functions (corresponding to rational numbers) and analytic functions (Section 3–16) (corresponding to real numbers).

A rational number (Section 1–8) may be defined as the indicated quotient of two integers a/b, where $b \neq 0$. A *rational function* of an indeterminate x may be defined as the indicated quotient of two polynomials $f(x)/d(x)$, where $d(x)$ is not identically zero. Similarly, if $f(x)$ and $d(x)$ are polynomials in a variable x, then the indicated quotient $f(x)/d(x)$ is called a rational function of the variable x and is defined for all values of x such that $d(x) \neq 0$. It is undefined when $d(x) = 0$, since division by zero is undefined. For example, $x^2 + x + 1$ is different from zero for all real values of x, and thus the rational function $(2x^2 - x + 1)/(x^2 + x + 1)$ is defined for all real values of x. The rational function $(x^2 - 1)/(x - 2)$ is defined when x is an indeterminate or when the variable x has a value different from 2. It is undefined when the variable x takes on the value 2.

Several of the above concepts may be restated using the terminology in Section 1–18. The ring of integers I has the field of rational numbers R as its quotient field. When the symbol x is adjoined to the field R, the ring $R[x]$ of polynomials in x with rational coefficients is obtained. The quotient field of $R[x]$ is $R(x)$, the field of rational functions of x. In general, if T is any integral domain, then $T[x]$ is the ring of polynomials in x with coefficients in T, and $T(x)$ denotes the quotient field of $T[x]$. This concept is important to us in that we shall at times consider a polynomial in x and y, such as

$$3x^2y + 4x - y^2 + 5,$$

as a polynomial in x with polynomials in y as coefficients, i.e., any polynomial $p(x, y)$ in $R[x, y]$ may be considered as a polynomial $p(x)$ in $T[x]$, where $T = R[y]$.

EXERCISES

In each exercise, indicate the values of the real variable x for which the rational function $y = f(x)/d(x)$ is defined.*

1. $3x + 2y = 1$
2. $y = \dfrac{x^2 - 4x + 3}{x^2 - 3x + 2}$
3. $xy = 5$
4. $x^2 + 2xy = y - 5$
5. $y = \dfrac{x^3 + 7x^2 + 1}{x^2 - 5x + 8}$

3–4 Divisibility. We now start a discussion of the ring of polynomials which is very similar to the discussion of the ring of integers in Chapter 2. Most of the definitions and theorems from Chapter 2 will be rephrased to apply to polynomials. This parallel development should give additional meaning both to the present discussion and to the theory of numbers.

A polynomial $d(x)$ *divides* a polynomial $f(x)$ if and only if there exists a polynomial $q(x)$ such that $f(x) = d(x) \cdot q(x)$ for all values of x. For example, $x - 1$ divides $x^2 - 1$; x divides $2x$; $2x - 2$ divides $3x^2 - 3$ in the ring of polynomials with rational coefficients but does not divide $3x^2 - 3$ in the ring of polynomials with integral coefficients, since the quotient $q(x)$ does not have integral coefficients in this case. The phrase "for all values of x" will be used throughout this text to indicate that a relation holds for all values of the variable x for which the expressions in the relation are defined. When the set of numbers from which the coefficients are taken is infinite, this phrase also indicates that the relation holds for any indeterminate x [7, 82]. Equations that hold for indeterminates are sometimes called *identities.*

Theorem 2–1 may now be stated for polynomials $f(x)$, $g(x)$, $d(x)$ as follows: If $d(x)$ divides $f(x)$ and $f(x)$ divides $g(x)$, then $d(x)$ divides $g(x)$. If $d(x)$ divides $f(x)$ and $d(x)$ divides $g(x)$, then $d(x)$ divides $f(x) + g(x)$ and $f(x) - g(x)$. The proof of this theorem is exactly analogous to that given for Theorem 2–1 (Exercise 9).

When the set of *allowable coefficients* (i.e., the set of numbers from which the coefficients of the polynomials under consideration may be chosen) forms a number system or a field, the only polynomials that divide every polynomial are the nonzero constants, i.e., the polynomials of degree zero. Thus the nonzero constants are the *units* for the ring of polynomials. However, only $+1$ is unity, the identity for multiplication.

* The zeros of a quadratic polynomial are discussed in Section 4–5.

In the theory of numbers (Theorem 2–8) we considered any integer different from zero as the product of a unit and a positive integer. In the theory of polynomials we define a polynomial with initial $+1$ to be a *monic polynomial.* Then, assuming that the set of allowable coefficients forms a number system, any polynomial (3–1) except the constant zero may be expressed as the product of a unit and a monic polynomial. For example, $2x^2 - 2 = 2(x^2 - 1)$ and $3x + 2 = 3(x + \frac{2}{3})$. The last example illustrates the necessity for the assumption that the set of allowable coefficients forms a number system, so that division by the initial of the polynomial will be possible.

The correspondence between monic polynomials and positive integers is also evident in the following definition of *the* greatest common divisor of two polynomials. Any polynomial $d(x)$ that divides both $f(x)$ and $g(x)$ is a common divisor of $f(x)$ and $g(x)$. If $d(x)$ is a monic polynomial and every common divisor of $f(x)$ and $g(x)$ divides $d(x)$, then $d(x)$ is *the greatest common divisor* of $f(x)$ and $g(x)$. Similarly, the definitions of common multiple, the least common multiple, and relatively prime are exactly analogous (Exercise 10) to those given for integers (Section 2–1). In the theory of polynomials, $2x$ and $2x^2 - 2$ are relatively prime, since their greatest common divisor is a unit.

Two integers have the same absolute value or numerical value if each may be expressed as the product of the other and a unit. Two polynomials are called *associates* if each may be expressed as the product of the other and a unit, that is, $f(x)$ and $g(x)$ are associates if $f(x)|g(x)$ and $g(x)|f(x)$. For example, $x - 2$, $5x - 10$, $7x - 14$, and $x/2 - 1$ are all associates when rational coefficients are allowed.

When the set of allowable coefficients forms a number system, every polynomial $p(x)$ except the constant zero has a unique associated monic polynomial. The units defined above are precisely the associates of unity. Two polynomials that are not associates are said to be *independent.*

EXERCISES

1. Does the set of polynomials of even degree in x form a ring? Does the set of polynomials in x^2 form a ring? Does either of these sets also form an integral domain?

2. List three polynomials and then express each as the product of a unit and a monic polynomial.

3. Can every polynomial in an arbitrary ring of polynomials be expressed as the product of a unit and a monic polynomial? Give examples.

4. Repeat Exercise 3 for an arbitrary integral domain.

5. Repeat Exercise 3 for polynomials with coefficients from an arbitrary field.

6. Give three associates of $x^3 - x^2 + 7x - 5$.

7. Give two independent polynomials having a common divisor $x - 2$.

8. Give two polynomials that are associates when the set of allowable coefficients is the set of real numbers but are not associates when the set of allowable coefficients is the ring of integers.

9. Prove Theorem 2–1 for polynomials in x with complex coefficients.

10. Define *common multiple*, *the least common multiple*, and *relatively prime* for polynomials in x.

3–5 Division Algorithm. In Section 2–2 the Division Algorithm was stated for integers as follows: If a and b are any two positive integers, there exist integers q and r, $0 \leqq q$, $0 \leqq r < a$, such that $b = qa + r$. In the theory of polynomials, the condition that the integers be positive could be replaced by the condition that the polynomials be monic polynomials. In this case we would have: If $p(x)$ and $q(x)$ are any two monic polynomials, there exist polynomials $s(x)$ and $r(x)$ such that $p(x) = s(x) \cdot q(x) + r(x)$ for all values of x, and either $r(x)$ is identically zero or the degree of $r(x)$ is less than that of $q(x)$. For example, if $p(x) = x^2 - 5x + 6$ and $q(x) = x - 3$, then $s(x) = x - 2$ and $r(x) = 0$; if $p(x) = x^3 - 2x^2 + 7x - 5$ and $q(x) = x^2 - x + 1$, then $s(x) = x - 1$ and $r(x) = 5x - 4$. This form of the Division Algorithm can be readily proved, but it is not the most useful form of the theorem. One disadvantage is that even when $r(x)$ is not identically zero, it is not necessarily a monic polynomial. For example, $r(x) = 5x - 4$ in the above example.

It is customary to replace the condition that the polynomials $p(x)$ and $q(x)$ be monic polynomials by an assumption, as in the discussion of monic polynomials (Section 3–4), that the set of allowable coefficients form a field such as the rational, real, or complex number system. Under this assumption the Division Algorithm may be stated for polynomials in one variable as follows:

THEOREM 3–1. *If $p(x)$ and $q(x)$ are any two polynomials with coefficients from a field, there exist polynomials $s(x)$ and $r(x)$ such that $p(x) = s(x) \cdot q(x) + r(x)$ for all values of x, and either $r(x)$ is identically zero or the degree of $r(x)$ is less than that of $q(x)$.*

The proof of Theorem 3–1 corresponds essentially to the short "proof" of the Division Algorithm in Section 2–2, since the properties of our number system have now been established. The details of the proof are left as an exercise for the reader. The necessity of the assumption that the set of allowable coefficients forms a field is evident from the following example: If $p(x) = x^6 - 2x^5 + 3x^2 + 1$ and $q(x) = 2x^2 + 3x + 1$, then we have

$$x^6 - 2x^5 + 3x^2 + 1 =$$

$$\left(\frac{x^4}{2} - \frac{7}{4}x^3 + \frac{19}{8}x^2 - \frac{43}{16}x + \frac{139}{32}\right)(2x^2 + 3x + 1) + \left(-\frac{331}{32}x - \frac{107}{32}\right),$$

where $$s(x) = \frac{x^4}{2} - \frac{7}{4}x^3 + \frac{19}{8}x^2 - \frac{43}{16}x + \frac{139}{32}$$

and $$r(x) = -\frac{331}{32}x - \frac{107}{32}.$$

Thus Theorem 3–1 involves only ring operations upon x, but all four rational operations upon the coefficients.

The fact that the Division Algorithm does not extend immediately to polynomials in two or more variables can be easily verified. Given two polynomials $p(x, y)$ and $q(x, y)$, we seek polynomials $s(x, y)$ and $r(x, y)$ such that

$$p(x, y) = q(x, y) \cdot s(x, y) + r(x, y) \tag{3–2}$$

for all values of x and y, and such that $r(x, y)$ is either identically zero or has degree less than that of $q(x, y)$. In particular, for $p(x, y) = x$ and $q(x, y) = y$, we seek polynomials $s(x, y)$ and $r(x, y)$ such that

$$x = y \cdot s(x, y) + r(x, y) \tag{3–3}$$

for all values of x and y, i.e., such that (3–3) is an identity (Section 3–4). Since $q(x, y) = y$ has degree one, $r(x, y)$ must be a constant. Since the degree of the right side of the identity (3–3) cannot exceed the degree of the left side, $s(x, y)$ must also be a constant. Thus we are seeking constants m and b such that $x = my + b$ for all values of x and y. Since (Section 3–1) there do not exist any constants m and b satisfying these conditions, it is not possible to find polynomials $s(x, y)$ and $r(x, y)$ satisfying (3–3). Thus, except for special cases, it is not possible to find polynomials $s(x, y)$ and $r(x, y)$ satisfying (3–2). Therefore Theorem 3–1 must be altered before it can be applied to polynomials in two or more variables. For example, (3–3) may be written in the form $x = 1 \cdot y + (x - y)$, where $r(x, y)$

$= x - y$ has the same degree as $y = q(x, y)$. A more useful modification is considered in Section 3–7.

Just as in the theory of numbers, the Division Algorithm for polynomials serves as the foundation for the Euclidean Algorithm for polynomials (Section 3–7). In view of the above difficulty with polynomials in two variables, we shall expect some modification of the Euclidean Algorithm to be necessary when polynomials in two or more variables are considered.

We shall apply the Division Algorithm for polynomials in one variable to the calculation of the greatest common divisor of two polynomials (Section 3–7), to the expression of a polynomial $p(x)$ in the form $q(bx + d)$ (Section 3–8), to the determination of the exact number of distinct real roots of a polynomial equation with real coefficients (Section 4–12), and to the determination of the multiple roots of a polynomial equation with complex (real or imaginary) coefficients (Section 4–13).

Exercises

Find $s(x)$ and $r(x)$ in Theorem 3–1 for each of the following pairs of polynomials:

1. $p(x) = x^2 - 3x + 4$, $q(x) = x - 2$.
2. $p(x) = 2x^2 - 3x + 4$, $q(x) = 3x - 2$.
3. $p(x) = x^3 - 5x^2 + 7x + 11$, $q(x) = x^2 + x - 1$.
4. $p(x) = x^4 + 3x^3 - 2x^2 + 2x - 1$, $q(x) = 3x^2 - 2x + 5$.
5. $p(x) = (1 - \sqrt{2})x^3 + (1 + \sqrt{2})x^2 + \sqrt{2}$, $q(x) = (1 + \sqrt{2})x^2 + (2 - \sqrt{2})$.

3–6 Irreducible polynomials. In Section 2–3 the integers were classified according to the integers that they divide, or are divisible by. It was found that every integer belongs to one of four classes: zero, units, prime numbers, composite numbers. Similarly, we shall find that every polynomial belongs to one of four classes: zero, units, irreducible polynomials, reducible polynomials.

The units divide every polynomial and, when the set of allowable coefficients forms a field, are composed of the nonzero constants (Section 3–4). A polynomial that is not zero or a unit is said to be *irreducible* if its only divisors are its associates and the units. A polynomial is called *reducible* if it has two or more irreducible divisors (not necessarily distinct). All linear polynomials are irreducible. The irreducibility of polynomials of degree greater than one often depends upon the set of allowable coefficients (Section 3–4). For example, $x^2 - 2$ is irreducible in the ring of polynomials with rational coefficients, reducible in the ring of polynomials with real coefficients.

In the theory of numbers it was convenient to assume that negative prime numbers were expressed as a product of a unit and a positive prime number. In the theory of polynomials it will often be convenient to assume that every irreducible polynomial is expressed as the product of a unit and an irreducible monic polynomial (Section 3–4).

We now use the above definitions and summarize some of the correspondences between the elements and properties of the theory of numbers and the theory of polynomials. When the set of allowable coefficients forms a field, the integers m with identity for addition, zero, and units $+1$ and -1 correspond to the polynomials $p(x)$ with identity for addition, zero, and units b, where b is any element different from zero in the set of allowable coefficients. Any integer different from zero may be expressed as the product of a unit and a positive integer; any polynomial that is not identically zero may be expressed as the product of a unit and a monic polynomial. Prime and composite integers correspond respectively to irreducible and reducible polynomials. The absolute value of an integer corresponds to the degree of a polynomial. For example, an integer m with $|m| = 0$ or a polynomial $p(x)$ without degree is identically zero; an integer m with $|m| = 1$ or a polynomial with degree zero is a unit; an integer m with $|m| > 1$ or a polynomial with positive degree is neither zero nor a unit. Most of these basic correspondences between the ring of integers m and the ring of polynomials $p(x)$ are given in the following array:

integers m	polynomials $p(x)$
zero	zero
unity, $+1$	unity, $+1$
units, $+1$ and -1	constants different from zero
positive integers	monic polynomials
prime integers	irreducible polynomials
composite integers	reducible polynomials
absolute value of m	degree of $p(x)$
$\|m\| > 1$	$p(x)$ has positive degree

These correspondences will be useful in restating for polynomials some of the theorems in the theory of numbers. For example, Theorem 2–2 may be stated for polynomials in the form:

THEOREM 3–2. *Every polynomial of positive degree has an irreducible monic polynomial divisor.*

The proof of this theorem may be obtained from that for Theorem 2–2 using the above correspondences. Let a polynomial $p(x)$ of degree m be given. If $p(x)$ is irreducible, it has its associated monic polynomial as an irreducible monic polynomial divisor. If $p(x)$ is not irreducible, then $p(x) = p_1(x) \cdot p_2(x)$, where $p_j(x)$ has positive degree m_j for $j = 1, 2$ and $m_1 + m_2 = m$. If no $p_j(x)$ is irreducible, then $p(x) = p_{11}(x) \cdot p_{12}(x) \cdot p_{21}(x) \cdot p_{22}(x)$, where no p_{jk} is a unit, that is, $0 < m_{jk}$. This process must terminate after a finite number of steps, since the sum of the positive integers m_{jk} is a given positive integer m (Exercise 7, Section 3–1). Thus $p(x)$ has at most m divisors and must have an irreducible polynomial divisor. When the set of allowable coefficients forms a field, every polynomial different from zero has a monic polynomial among its associates. Then every polynomial $p(x)$ of nonnegative degree has an irreducible monic polynomial divisor.

Theorem 2–3 is restated for polynomials in Exercise 1. Theorem 2–4 does not immediately extend to the theory of polynomials. For example, in the ring of polynomials with real coefficients the set of irreducible polynomials has an uncountably infinite subset, since $x - b$ is irreducible for every real number b. Most of the other theorems of Sections 2–3 and 2–4 are restated for polynomials in the following exercises. The proofs may be obtained from those in Chapter 2, using the correspondences listed above.

Many of the following exercises and their proofs may be restated for polynomials in several variables. For example (Exercise 9), the Unique Factorization Theorem (Theorem 2–8), holds for polynomials in any finite number of variables with integral, rational, real, or complex coefficients [7; 97–100].

EXERCISES

1. Prove that any reducible polynomial $p(x)$ of degree m has an irreducible monic polynomial divisor of degree $\leqq m/2$.
2. Give examples for Exercise 1 when $m = 2, 3, 5$, and 7.
3. Prove that if $p(x)$ is an irreducible polynomial and $q(x)$ is any polynomial, then either $p(x)$ divides $q(x)$ or they are relatively prime.
4. Give examples illustrating both cases in Exercise 3.
5. If $r(x)$ and $s(x)$ are two polynomials each of degree less than m, and $p(x)$ is an irreducible polynomial of degree m, prove that $p(x)$ does not divide $r(x) \cdot s(x)$.
6. Give two examples illustrating Exercise 5.
7. If an irreducible polynomial $p(x)$ divides the product

$$q_1(x) \cdot q_2(x) \cdot \ldots \cdot q_n(x),$$

prove that $p(x)$ divides at least one of the polynomials $q_1(x), q_2(x), \ldots, q_n(x)$, where n is any positive integer.
8. Give two examples illustrating Exercise 7.
9. Prove that except for the order of the factors every polynomial that is not identically zero can be represented in one and only one way as a product of a unit and a finite number of irreducible monic polynomials.
10. Prove that the number of independent divisors of

$$p(x) = e[r_1(x)]^{a_1} \cdot [r_2(x)]^{a_2} \cdot \ldots \cdot [r_k(x)]^{a_k}$$

is $(a_1 + 1)(a_2 + 1) \cdots (a_k + 1)$, where the $r_j(x)$ are distinct irreducible polynomials, e is a constant, and the a_j are positive integers.
11. Give an example illustrating Exercise 10 where $k = 3$, and list the independent divisors.
12. Repeat Exercise 11 where $e \neq 1$ and $k > 3$.
13. If $r(x)$ is the greatest common divisor and $s(x)$ is the least common multiple of two monic polynomials $p(x)$ and $q(x)$, prove that $r(x) \cdot s(x) = p(x) \cdot q(x)$.
14. Restate the Postulate of Archimedes (Section 2–2) for polynomials and give an example.
15. Restate for polynomials each of the three parts of Theorem 2–9.
16. Give examples illustrating each statement in Exercise 15.

3–7 Euclidean Algorithm. The Euclidean Algorithm was used in Section 2–5 to find the greatest common divisor of two integers without expressing either integer in terms of its prime factors. We now use the same procedure for polynomials $p(x)$ to find the greatest common divisor of two polynomials without expressing either polynomial in terms of its irreducible factors. Theorem 2–11 may be restated for polynomials as follows:

THEOREM 3–3. *The greatest common divisor $r_n(x)$ of any two polynomials $f_0(x)$ and $f_1(x)$ of positive degree and with coefficients from a field can be found as the last nonvanishing polynomial remainder in the Euclidean Algorithm. There exist polynomials $A(x)$ and $B(x)$ such that $r_n(x) = A(x) \cdot f_0(x) + B(x) \cdot f_1(x)$ for all values of x.*

We may prove the first part of the theorem by repeated applications of the Division Algorithm (Section 3–5). The procedure for the polynomials $f_0(x)$ and $f_1(x)$ is illustrated by the following array, where b is a nonzero constant to be determined, the degree of $r_1(x)$ is less than that of $f_1(x)$, and the degree of $r_j(x)$ is less than that of $r_{j-1}(x)$ for $j = 2, 3, \ldots, n$.

$$\begin{aligned} f_0(x) &= q_1(x)f_1(x) + r_1(x), \\ f_1(x) &= q_2(x)r_1(x) + r_2(x), \\ r_1(x) &= q_3(x)r_2(x) + r_3(x), \\ &\cdot \\ &\cdot \\ &\cdot \\ r_{n-2}(x) &= q_n(x)r_{n-1}(x) + br_n(x), \\ r_{n-1}(x) &= q_{n+1}(x)r_n(x). \end{aligned}$$

In this array it is convenient to assume that the initial (Section 3–1) of $r_n(x)$ has been made $+1$ by dividing both sides of the next to the last equation by a suitable constant b. Then $r_n(x)$ is a monic polynomial and, using the same reasoning as in Section 2–5, is the greatest common divisor of $f_0(x)$ and $f_1(x)$. With the convention that $r_n(x)$ is to be a monic polynomial, constant factors (units) may be inserted or removed from any equation of the array. Essentially, this means that any polynomial remainder may be replaced by any one of its associates at any time.

The proof of the existence of polynomials $A(x)$ and $B(x)$ in the second part of the theorem is also similar to that in Section 2–5. We successively express $r_1(x), r_2(x), \ldots, r_n(x)$ in the form $r_j(x) = A_j(x)f_0(x) + B_j(x)f_1(x)$. All further details of the proof are left as an exercise for the reader.

The Euclidean Algorithm, like the Division Algorithm on which it is based, does not extend directly to polynomials in two or more unknowns. For example, even though the greatest common divisor of x and y is 1, there does not exist an equation of the form $Ax + By = 1$, where A and B are constants, that holds for all values of x and y. An extension of the theorem is possible when $f_0(x_1, x_2, \ldots, x_n)$ and $f_1(x_1, x_2, \ldots, x_n)$ are considered as polynomials in x_n with polynomials in the first $n - 1$ variables as coefficients. The A and B are then rational functions of the first $n - 1$ variables, since all four rational operations are used upon the coefficients in the Division Algorithm (Section 3–5). For example, considering the polynomials x and y as polynomials in y with coefficients in x, we have $(1/x) \cdot x + 0 \cdot y = 1$.

Arrays similar to those used in Section 2–5 to find n_k, A, B may also be used for polynomials in one variable. In general, we have

f_0	f_1	r_1	r_2	$\ldots$	r_n	0
	q_1	q_2	q_3	$\ldots$	q_{n+1}	

Since the calculation of the q's and r's now usually requires written computations, we interchange the order of the two rows and use the array

	q_1	q_2	q_3	$\ldots$	q_{n+1}	
f_0	f_1	r_1	r_2	$\ldots$	r_n	0
q_1f_1	q_2r_1	q_3r_2	$\ldots$			
a_1r_1	a_2r_2	a_3r_3	$\ldots$			

where $a_1r_1 = f_0 - q_1f_1$, $a_2r_2 = f_1 - q_2r_1$, ... and the a_j's are arbitrary constants.

If $f_0(x) = x^4 - 3x^3 + 2x$ and $f_1(x) = x^3 - x$, we have the array

	$x - 3$	$x + 1$
$x^4 - 3x^3 + 2x$	$x^3 - x$	$x^2 - x$ 0
$x^4 - x^2$	$x^3 - x^2$	
$-3x^3 + x^2 + 2x$	$x^2 - x$	
$-3x^3 + 3x$	$x^2 - x$	
$x^2 - x$	0	

whence the greatest common divisor is $x^2 - x$. This array can be extended to give $A(x)$ and $B(x)$, as in Section 2–5. The algebraic details in the calculation of $r_n(x)$ may often be noticeably simplified by the use of cross multiplication [37].

As in the numerical case, the greatest common divisor could be obtained by inspection if both polynomials were expressed in terms of their irreducible factors (Exercise 9, Section 3–6). The Euclidean Algorithm gives a procedure for determining the greatest common divisor of two polynomials without factoring the polynomials. This algorithm also has very practical and important applications in the determination of the number of roots of a polynomial equation (Sections 4–12 and 4–13).

In the next section we continue our development of the theory of polynomials and seek a concept corresponding to that of bases (Section 2–6) in the theory of numbers.

Exercises

Use the Euclidean Algorithm to find the greatest common divisor of each of the following pairs of polynomials.

1. $x^2 - 5x + 6$ and $x^2 - 4$.
2. $x^3 - 3x^2 + 3x - 1$ and $x^2 - 2x + 1$.
3. $x^3 + 2x + 20$ and $3x^2 + 2$.
4. $x^4 - 6x^3 + 7x^2 + 6x - 2$ and $2x^3 - 9x^2 + 7x + 3$.

3–8 Change of variable. Any positive integer m may be expressed as a polynomial in an arbitrary positive integer $n > 1$ with coefficients from the set of integers $0, 1, 2, \ldots, n - 1$ (Theorem 2–13). There are several possible statements for polynomials corresponding to this theorem. The following is one of the most useful.

THEOREM 3–4. *Any polynomial $p(x)$ with coefficients from a field may be expressed as a polynomial in an arbitrary linear polynomial $bx + d$.*

Given a polynomial $p(x) = x^3 - 6x^2 + 2$, we may designate its zeros as r, s, t and seek a new cubic $q(y)$ with zeros $r - 2$, $s - 2$, $t - 2$. This process of reducing the zeros of a polynomial is frequently used in solving cubic equations (Section 4–9). Methods for obtaining the new polynomial will be considered in Sections 3–15 and 4–3. In the above case, it would then be found that $q(y) = y^3 - 12y - 14$ or $q(x - 2) = (x - 2)^3 - 12(x - 2) - 14$, where $x - 2 = y$. Theorem 3–4 states that for any numbers $b \neq 0$ and d in the set of allowable coefficients, any polynomial $p(x)$ may be written in the form $q(bx + d)$.

The following proof of Theorem 3–4 corresponds closely to that for Theorem 2–13. Given a polynomial $p(x)$ of degree m and a linear polynomial $bx + d$, we apply Theorem 3–1 to obtain

$$p(x) = p_1(x) \cdot (bx + d) + r_1,$$

where r_1 is a constant since, if it is different from zero, its degree must be less than that of $bx + d$. Furthermore, the degree of $p_1(x)$ is one less than that of $p(x)$. If $p_1(x)$ has positive degree, we may repeat the process. In general, since $p(x)$ has degree m, we repeat the process m times and obtain a sequence

$$\begin{aligned} p(x) &= p_1(x)(bx + d) + r_1, \\ p_1(x) &= p_2(x)(bx + d) + r_2, \\ &\cdot \\ &\cdot \\ &\cdot \\ p_{m-2}(x) &= p_{m-1}(x)(bx + d) + r_{m-1}, \\ p_{m-1}(x) &= p_m(x)(bx + d) + r_m, \end{aligned}$$

where the p_i's are polynomials of degree $m - j$ and the r's are constants. Let the constant $p_m(x)$ be designated by r_{m+1}. The above equations may then be used exactly as in Section 2–6 to obtain

$$\begin{aligned} p(x) &= r_{m+1}(bx + d)^m + r_m(bx + d)^{m-1} + \cdots + r_2(bx + d) + r_1 \\ &= q(bx + d). \end{aligned}$$

Synthetic division (Section 4–2) and Taylor's Theorem (Section 3–15) are very useful in performing the computations that are frequently necessary in order to apply Theorem 3–4. Accordingly, we shall not consider detailed applications of Theorem 3–4 until after these other concepts have been introduced.

Most of the remaining topics that we considered in the theory of numbers (congruence, residue class, ϕ-function, Diophantine problem) have one or more interpretations in the theory of polynomials. However, since these interpretations will not be needed in our discussion of the theory of equations, we shall mention only the concept of an ideal corresponding to a congruence (Section 3–9). Then we shall reorient our development of the theory of polynomials. In the first eight sections of this chapter, we have developed correspondences between the ring of integers and the ring of polynomials. In Section 3–10 we shall start the development of several types of functions from the ring of polynomials corresponding to the development in Chapter 1 of several number systems from the ring of integers. Zeros of polynomials will be considered in Chapter 4.

EXERCISES

1. Write $p(x) = x^2 + 3x - 1$ as $q(x + 2)$.
2. Write $p(x) = x^3 + 3x^2 + 5$ as $q(x + 1)$.
3. Write $p(x) = x^4 + 8x^3 - 7x + 11$ as $q(x - 2)$.

***3–9 Ideals.** We shall conclude our development of the correspondences between the theory of polynomials and the theory of numbers in Chapter 2 with a brief mention of a concept corresponding to congruences (Section 2–8).

Two integers a and b are congruent modulo an integer m if their difference is divisible by m. Two polynomials $p(x)$ and $q(x)$ are congruent modulo a polynomial $m(x)$ if their difference is divisible by $m(x)$. The set of all integral multiples of an integer m forms a residue class [0] (mod m) and constitutes the *ideal* of m in the ring of integers. The set of all polynomial multiples of $m(x)$ constitutes

the ideal of $m(x)$ in the ring of polynomials. For example, $x^2 + 2x$, x^5, $x^{17} - 12x^2$ are each in the ideal of x, whereas $x^2 + b$ with $b \neq 0$ is not in the ideal of x.

Residue classes may be defined in terms of the ideal of $m(x)$. However, the number of residue classes is no longer finite. For example, in the ring of polynomials with integral coefficients, each integer represents a distinct residue class with respect to the ideal of x. The analogs, when such exist, for most of the theorems on residues in the theory of numbers are beyond the scope of our brief discussion. An excellent introduction to this subject may be found in [34].

Exercises

1. In the ring of polynomials with integral coefficients, describe the ideal of (a) x, (b) x^2, (c) $x + 1$.

2. A subset S of one or more elements from a ring R is an ideal if (i) the difference of any two elements of S is an element of S, and (ii) the product of any element of S by an element of R is an element of S. This statement is commonly used to define an ideal. Show that each ideal in Exercise 1 satisfies this definition.

3. Use the definition in Exercise 2 and prove that any ideal S in a ring R must also be a subring of R.

4. Indicate which of the following are ideals: (a) the ring of even integers in the ring of integers, (b) the ring of integers in the ring of rational numbers.

5. An ideal S in a ring R is called a *principal ideal* if every element of S is of the form rb, where b is a fixed element and r is an arbitrary element of R. Each of the ideals in Exercise 1 is a principal ideal; in fact, whenever the Euclidean Algorithm holds in a ring, every ideal in that ring is a principal ideal. Give an example of an ideal that is not a principal ideal.

3–10 Functions. Given any polynomial $p(x)$, we may associate a unique number $p(b)$ with each numerical value b assigned to x. Thus the numerical values assumed by the polynomial $p(x)$ depend upon the set of values S assumed by the variable x. This notion of dependence is the basis for the following definition of function.

The variable y is said to be a *function* of the variable x over a set of numbers S if corresponding to each value of x from S there are one or more values of y. The variable x is called the *independent variable;* y, the *dependent variable.* Formally, the function is the rule or channel by means of which values of x give rise to values of $f(x)$. This function or rule may be expressed as a polynomial in x, as a graph in the xy-plane, and in many other ways. The functional relation between x and y is indicated in symbols by $y = f(x)$. The set of values S is called the domain of definition or, briefly, the *domain* of f. The set of values assumed by y is called the range of values or *range* of f. If there is exactly one value of y corresponding to each value of x from S, y is said to be a *single-valued function* of x over S. If there are two or more values of y for each x, y is a *multiple-valued function* of x.

In terms of these definitions, the statements at the beginning of this section imply that any polynomial $p(x)$ is a single-valued function of the variable x. This statement is a consequence of the definitions of Chapter 1, since for any given numerical value, say b of x, $p(b)$ involves only a finite number of the operations of addition, subtraction, and multiplication of numbers. Each of these operations has been defined so as to be unique. Consider, for example, $p(x) = x^3 - 7x^2 + 3x + 2$ at $x = 5$, where

$$p(5) = 5 \cdot 5 \cdot 5 - 7 \cdot 5 \cdot 5 + 3 \cdot 5 + 2$$

is uniquely defined.

We have seen that the set of values, or range of the function, is dependent upon the set of values S assumed by the variable x, i.e., the domain of the function. It is convenient to identify the sets S to which x is commonly restricted. The set of values S often consists of one or more intervals where the set of all real numbers satisfying any one of the following relations is called an *interval:* $x < a$, $x \leqq a$, $a < x < b$, $a \leqq x < b$, $a \leqq x \leqq b$, $a < x \leqq b$, $b \leqq x$, $b < x$ for any real numbers a, b. The set $a < x < b$ is also called a *segment* or *open interval;* $a \leqq x < b$ is neither open nor closed and is sometimes called a *semi-closed interval;* $a \leqq x \leqq b$ is called a *closed interval.* When the set S consists of all real numbers or a single interval of real numbers such as $x < 0$, $0 < x < 1$, or $2 \leqq x$, x is called a *continuous real variable.* When the set S consists of all positive integers, x is called a *positive integral variable.*

We now prepare to define a continuous function (Section 3–12). This preparation is one of the main purposes of the present and the following section. Continuity is an important property of all polynomials in continuous variables. In fact, we shall find (Exercise 4, Section 3–13) that when x is a continuous real variable, the polynomial $p(x)$ is a continuous function of x. This property of polynomials is a fundamental one and may be used to locate roots of any polynomial $p(x)$ (Exercise 5, Section 3–13; Section 4–5). One of the best definitions of a continuous function involves limits.

The concept of limit is usually considered to be a topic of analysis where the three main subdivisions of mathematics are algebra, geometry, and analysis. However, the fundamental concepts of algebra cannot be compartmentalized into a closely knit entity entirely separate from geometry and analysis. A postulate for the existence of all real numbers may also be used to postulate continuity on a line in geometry (Section 1–12). We have used geometric representations of algebraic relations to clarify the concepts involved (Sections 1–12 and 1–16). Thus we now consider a few topics of analysis [limit (Section 3–11), continuity (Sections 3–12 and 3–13), and derivative (Section 3–14)] that will be useful in our discussion of the theory of equations (Chapter 4) and in our development in this chapter of the following correspondences between numbers and functions:

integers	polynomials
rational numbers	rational functions
algebraic numbers	algebraic functions
transcendental numbers	transcendental functions
real numbers	analytic functions

Exercises

1. Indicate which of the following functions are single-valued functions of x (consider only real values of x and y):

(a) $y = x^3 - 3x + 1$
(b) $y = \sqrt{x^2 - 9}$
(c) $y = (x^2 - 9)^2$
(d) $y = \dfrac{x^2 + 1}{x^2 - 1}$
*(e) $y = 2^x$
*(f) $y = \log x$
*(g) $y = \sin x$
*(h) $y = \text{arc sin } x$

2. Give the domain of definition and range of values for each of the functions in Exercise 1.

3. List five multiple-valued functions of x.

4. Give three examples of each of the following types of intervals: (a) open, (b) closed, (c) neither open nor closed.

5. Indicate which of the following define functions of the real variable x:
(a) $y = x^3 - 3x + 1$,
(b) $y = x$ when $x > 0$, $-x$ when $x \leqq 0$,
(c) $y = x - [x]$, where $[x]$ indicates the greatest integer $\leqq x$,
(d) $y = x/(x^2 - 2)$,

* The asterisk indicates that the exercise involves concepts that have not been discussed in the present text but that should be familiar to most readers.

(e) $y = 1$ when x is rational, 0 when x is irrational,
*(f) $y = \tan x$,
(g) $y = 2$ when $x < -1$, $-x$ when $-1 \leqq x \leqq 0$, $1/x$ when $0 < x$.

6. Repeat Exercise 5 when the domain of definition of x is the set of (a) rational numbers, (b) positive integers.

7. If $f(y)$ is given by $x = y^n$, then the principal value of the inverse function $f^{-1}(x)$ is given by $y = x^{1/n}$. This relation may be used to define $x^{1/n}$ for (i) $x > 0$ and any integer $n \neq 0$, and (ii) any real value of x and any odd integer n (Exercise 13, Section 1–12). Designate the inverse functions of each of the following:

(a) $x = y + 2$
(b) $x = 2y$
(c) $x = 3y + 5$
(d) $x = y^3$
(e) $x = y^{1/3}$
(f) $x = (y + 1)/(y - 1)$
(g) $x = (ay + b)/(cy + d)$, where $ad - bc \neq 0$ (See Theorem 4–8)
(h) $y = g(x)$
*(i) $x = \sin y$
*(j) $x = 2^y$
*(k) $x = \log_3 y$.

8. Discuss the relations among the domains and ranges of $f(y)$ and $f^{-1}(x)$ in each part of Exercise 7.

9. Find the inverse functions of each of the following:

(a) $x = y + b$
(b) $x = by$
(c) $x = cy + b$
(d) $x = cy^n + b$
*(e) $x = a^y$, where $0 < a$
*(f) $x = \log_b y$, where $0 < b$

10. Use the definition of $x^{1/n}$ given in Exercise 7 and define $x^{m/n}$ for $x > 0$ and any integers m and n.

11. A real single-valued function $f(x)$ is an *increasing function* of x on an interval $a < x < b$ if $f(x + h) - f(x) > 0$ for every x and every h such that $a < x < x + h < b$. Prove that x^2 is an increasing function of x for $x > 0$.

12. Prove that x^n is an increasing function of x for $x > 0$ and any positive integer n. (*Hint:* Let $x + h = y$ and use Exercise 7, Section 1–4.)

13. Define a decreasing function of x and prove that x^{-n} is a decreasing function of x for any positive integer n and $0 < x < 1$.

14. Prove that a^x is an increasing function of the positive integral variable x when $a > 1$.

15. Repeat Exercise 12 where n may be any integer.

16. Prove that a^x is a decreasing function of the integral variable x when $0 < a < 1$.

17. Indicate the functions in Exercise 7 that are increasing functions of the real variable y. Specify intervals when necessary.

18. Indicate the increasing functions in Exercise 9.

* The asterisk indicates that the exercise involves concepts that have not been discussed in the present text but that should be familiar to most readers.

3-11 Limits. We first define the limit of an ordered set of real numbers

$$\{a_n\} = a_1, a_2, \ldots, a_n, \ldots.$$

The notation $\{a_n\}$ indicates that there exists a number a_n corresponding to each positive integer n, and the numbers a_n (not necessarily distinct) are considered in the order of their subscripts. For example, if $a_n = 1/n^2$, we have $a_1 = 1$, $a_2 = \frac{1}{4}$, $a_3 = \frac{1}{9}$, ... Such ordered sets are called *sequences* of numbers. We may also consider sequences of functions, such as $\{x^n - 1\}$.

The sequences

$$1, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{4}, \ldots, 1/n, \ldots,$$
$$-\tfrac{1}{2}, +\tfrac{1}{4}, -\tfrac{1}{8}, +\tfrac{1}{16}, \ldots, (-1)^n/2^n, \ldots;$$
$$.1, .01, .001, \ldots, 10^{-n}, \ldots$$

clearly have a common property in that for each sequence $\{a_n\}$ the term a_n can be chosen to have arbitrarily small absolute value by choosing n sufficiently large. In other words, $|a_n - 0|$ can be made less than any given positive number ϵ by a suitable choice of n. We let N denote a particular value of n such that $|a_n - 0| < \epsilon$ for all $n \geqq N$. Since $|a_n| > |a_{n+1}|$ in each of the given sequences, it is clear that $|a_n - 0| < \epsilon$ for $n = N$ implies that the same relation is also satisfied for any $n \geqq N$. In the terminology of analysis, we indicate the dependence of N upon ϵ by writing N_ϵ. For example, if $\epsilon = 1$ and $\{a_n\} = \{1/n\}$, then we may choose $N_1 = 2$. If $\epsilon = .01$ for the same sequence, we choose $N_{.01} = 101$. The necessity for this dependence of N upon the given ϵ is evident from the fact that for any given value of N one may often choose an ϵ such that the given N does not satisfy the desired conditions. In particular, if $\{a_n\} = \{1/n\}$ and $N = 1{,}658{,}972$ is given, one need only take $\epsilon = 10^{-7}$ to show that ϵ must be given first and N_ϵ chosen for the given ϵ. Using this terminology, we may describe the common property of the above sequences by saying that given any positive number ϵ there exists an integer N_ϵ such that $|a_n - 0| < \epsilon$ for all $n \geqq N_\epsilon$. Sequences with this property are said to approach zero as a limit and hence are called *null sequences.*

If we consider the sequences

$$.9, .99, .999, \ldots, (10^n - 1)/10^n, \ldots,$$
$$\tfrac{2}{1}, \tfrac{3}{2}, \tfrac{4}{3}, \ldots, (n+1)/n, \ldots,$$

we find that in each sequence the terms a_n approach 1 as n increases. Given $\epsilon = .1$, we may take $N_\epsilon = 2$ in the first sequence and $N_\epsilon = 11$ in the second sequence in order to have $|a_n - 1| < \epsilon = .1$ for all $n \geqq N_\epsilon$. Similarly, if $\epsilon = .001$, we may take $N_\epsilon = 4$ in the first sequence and $N_\epsilon = 1001$ in the second sequence. In general, *a sequence of numbers*

$$a_1, a_2, \ldots, a_n, \ldots$$

is said to approach a finite limit A, $\lim_{n\to\infty} a_n = A$, *if for every positive number* ϵ *there exists an integer* N_ϵ *such that for any* $n \geqq N_\epsilon$ *the inequality* $|a_n - A| < \epsilon$ *is satisfied.*

Many sequences such as $\{n\}$ and $\{-n^2\}$ may be defined to approach infinite limits (Exercise 8) [12; 33]. Many other sequences, such as $\{(-1)^n\}$ and $\{n(-1)^n\}$, oscillate and do not approach a limit of any kind. We do not consider such sequences, since we shall need to use only *convergent sequences,* i.e., sequences which approach a finite limit. A good general discussion of sequences may be found in [12; 27–41.]

The manipulations of numerical inequalities used in the determination of N_ϵ for a given sequence are based upon the definitions in Sections 1–6, 1–9, and 1–12. Briefly, given $a < b$, we have for any real number c

$$a + c < b + c,$$

and therefore

$$a - c < b - c.$$

Also

$$ac < bc \quad \text{if } c \text{ is positive,}$$
$$ac > bc \quad \text{if } c \text{ is negative.}$$

The convergence of a sequence may also be considered in terms of Cauchy sequences. A sequence of real numbers $\{a_n\}$ is called a *Cauchy sequence* if for any given positive number ϵ there exists an integer N_ϵ such that $|a_n - a_{n+k}| < \epsilon$ for $n \geqq N_\epsilon$ and all positive integers k. The Cauchy convergence criterion (Exercise 4) then states that every Cauchy sequence is a convergent sequence and, conversely, every convergent sequence is a Cauchy sequence. Thus a sequence may be proved to converge by proving that it is a Cauchy sequence.

All the above sequences were obtained by expressing a_n as a function of the positive integral variable n and considering the sequence of values of a_n as $n = 1, 2, \ldots$. We may also obtain sequences by expressing the general term, say a_x, as a function of a

continuous real variable x and considering the sequence of values of a_x corresponding to any sequence of real numbers taken as values of x. For example, let $a_x = x + 5$ and let x take on the values $1, 1/\sqrt{2}, 1/\sqrt{3}, \ldots, 1/\sqrt{n}, \ldots$. This concept will be used in the next two sections in the definition of a continuous function of a continuous variable.

Exercises

1. Prove that the following sequences are null sequences:

$$1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \ldots, \frac{1}{n}, \ldots,$$

$$\frac{1}{2}, -\frac{1}{4}, \frac{1}{8}, -\frac{1}{16}, \frac{1}{32}, \ldots, \frac{(-1)^{n+1}}{2^n}, \ldots,$$

$$.1, .01, .001, .0001, \ldots, \frac{1}{10^n}, \ldots$$

2. Write the first five terms and find the limit of each of the following sequences.

(a) $\{n^{-2}\}$, (b) $\{[5n + (-1)^{n+1}]/n\}$, (c) $\{(n^2 - 1)/n^2\}$.

3. Find an N_ϵ for each of the sequences in Exercise 2 (a) when ϵ is an arbitrary positive number, (b) when $\epsilon = .01$.

4. Prove the *Cauchy convergence criterion:* A sequence of real numbers $\{a_n\}$ approaches a finite limit A if and only if for any given positive number ϵ there exists an integer N_ϵ such that $|a_n - a_{n+k}| < \epsilon$ for $n \geqq N_\epsilon$ and all positive integers k [21; 35–36].

5. If $\lim_{n\to\infty} a_n = A$ and $\lim_{n\to\infty} b_n = B$, where A and B are real numbers, prove that

(a) $\lim_{n\to\infty} (a_n + b_n) = A + B$,

(b) $\lim_{n\to\infty} (a_n - b_n) = A - B$,

(c) $\lim_{n\to\infty} a_n b_n = AB$, and

(d) when $B \neq 0$, $\lim_{n\to\infty} \frac{a_n}{b_n} = \frac{A}{B}$.

6. The assumption that every convergent sequence of rational numbers has a limit is often used to postulate the existence of the real numbers. In Exercise 5 it has been indicated that limits of sequences of numbers may be added, subtracted, multiplied, and divided the same as numbers. Show that the order relations for limits of sequences are similar to but not exactly the same as those for numbers by proving that

(a) if $a_n \geqq 0$ and $\lim_{n\to\infty} a_n = A$, then $A \geqq 0$,

(b) if $a_n \geqq b_n$, $\lim_{n\to\infty} a_n = A$, $\lim_{n\to\infty} b_n = B$, then $A \geqq B$,

(c) it is possible to have $a_n > b_n$, $\lim_{n\to\infty} a_n = A$, $\lim_{n\to\infty} b_n = B$, and $A = B$.

7. Use Exercise 10, Section 3–10, and the definition of a real number as the limit of a sequence of rational numbers to define x^b for any positive real number b.

8. We may define the symbols $\lim_{n\to\infty} a_n = +\infty$, $\lim_{n\to\infty} -a_n = -\infty$ if for any $\epsilon > 0$ there exists an integer N_ϵ such that $a_n > 1/\epsilon$ for $n \geqq N_\epsilon$. The sequence $\{a_n\}$ is thus said to become positively infinite if for any given positive integer p we have $a_n > p$ for all sufficiently large values of n. Prove that if $\lim_{n\to\infty} a_n = +\infty$, $\lim_{n\to\infty} b_n = B$, where B is finite and $\lim_{n\to\infty} c_n = 0$, then

(a) $\lim_{n\to\infty} (a_n \pm b_n) = +\infty$,

(b) $\lim_{n\to\infty} (b_n - a_n) = -\infty$,

(c) $\lim_{n\to\infty} \frac{b_n}{a_n} = 0$,

(d) $\lim_{n\to\infty} a_n b_n = +\infty$ if $0 < B$,

(e) $\lim_{n\to\infty} \frac{b_n}{|c_n|} = +\infty$ if $0 < B$.

9. Use the results of Exercise 8 to justify the following conventions, where B is finite:

(a) $\infty \pm B = \infty$,
(b) $B - \infty = -\infty$,
(c) $B/\infty = 0$,
(d) $\infty \cdot B = \infty$ if $0 < B$,
(e) $B/0 = \infty$ if $0 < B$.

Note that there are also some undefined operations involving the symbols ∞ and 0: $\infty - \infty$, ∞/∞, $0/0$, $\infty \cdot 0$, 0^0, ∞^0, 1^∞.

10. Given an infinite sequence $\{a_n\}$ of real numbers, where n is a positive integral variable, the indicated sum $\sum_{n=1}^{\infty} a_n$ is called an *infinite series* of real numbers. The series *converges* to (has sum) S if and only if the sequence of partial sums, s_n, where $s_n = a_1 + a_2 + \cdots + a_n$, has limit S and S is finite. The series is said to *diverge* in all other cases, i.e., when (i) $\{s_n\}$ becomes positively infinite, (ii) $\{s_n\}$ becomes negatively infinite, (iii) $\{s_n\}$ oscillates and therefore S does not exist. Give examples illustrating each of these three types of divergent series.

11. Prove that if $\sum_{n=1}^{\infty} a_n = S$, $\sum_{n=1}^{\infty} b_n = R$, and C is any real number, then

(a) $\sum_{n=k}^{\infty} a_n = S - a_1 - a_2 - \cdots - a_{k-1}$,

(b) $C + \sum_{n=1}^{\infty} a_n = C + S$,

(c) $\sum_{n=1}^{\infty} Ca_n = CS$,

(d) $\lim_{n\to\infty} a_n = 0$,

(e) $\sum_{n=1}^{\infty} (a_n + b_n) = S + R$.

12. Restate each part of Exercise 11 in words and indicate its significance.

3–12 Continuity. One frequently hears it said that a continuous curve is one that can be drawn without lifting the pencil. All such

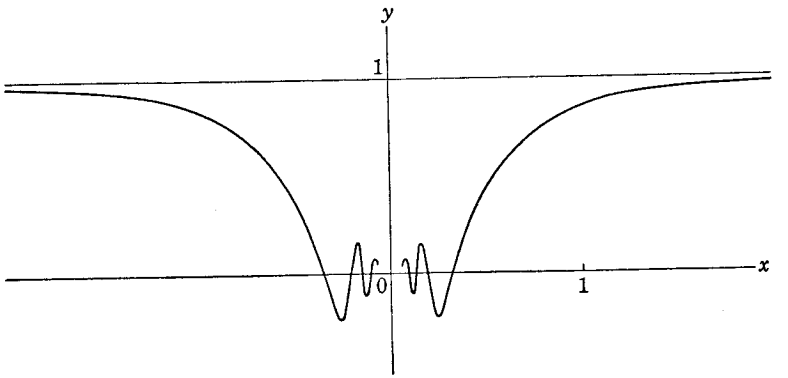


FIG. 3–1

curves are continuous, and the definition is easy to visualize. However, such a definition is not exact, since there do exist curves, such as $y = x \sin (1/x)$ in the neighborhood of the origin, that oscillate so rapidly that they cannot be drawn with a pencil (Fig. 3–1), and yet some of these curves are also continuous. A continuous curve is the graph of a continuous function and, conversely, the graph of a continuous function is a continuous curve. We shall use this fact to define continuity in terms of limits for both curves and functions.

Consider the function $y = f(x)$ graphed in Fig. 3–2. It is defined by

$$\begin{aligned} y &= 2 && \text{when } x < -1, \\ &= -x && \text{when } -1 \leqq x \leqq 0, \text{ and} \\ &= 1/x && \text{when } 0 < x. \end{aligned}$$

In this way we may associate exactly one value of y with each real value of x for all real numbers x. Thus (Section 3–10) we have a single-valued function $y = f(x)$ of a continuous real variable x. From the graph it would appear that the function is continuous for positive x but not for all x. Let us now examine the curve in more detail and find a basis for describing a curve as continuous.

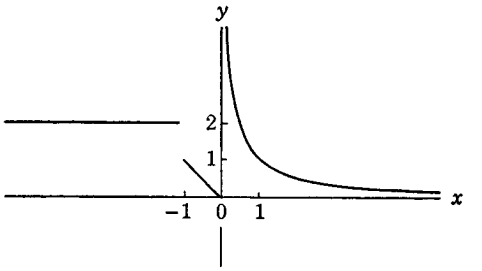


FIG. 3–2

Each break in the curve must occur at some point such as $x = -1$ and $x = 0$ in Fig. 3–2. Thus we shall be primarily concerned with continuity at a point. If a curve is continuous for every point of an interval (Section 3–10), it is defined to be continuous on that interval. The above curve appears to be, and is, continuous on each of the three segments $x < -1$, $-1 < x < 0$, $0 < x$. The points at which the curve jumps or breaks are called points of *discontinuity*. These points may be exactly defined in terms of the limits of sequences (Section 3–13).

Suppose $y = f(x)$ is defined as for Fig. 3–2. For any sequence of positive numbers $\{a_n\}$ having limit 2 there exists a corresponding sequence $\{1/a_n\}$ of values of $y = f(x)$. The function $f(x)$ is said to be continuous at $x = 2$, since for *every* sequence of values of x having the limit 2 the corresponding sequence of values of $f(x)$ has the limit $\frac{1}{2}$ and $f(2) = \frac{1}{2}$. The function $f(x)$ is discontinuous at $x = 0$, since there exist null sequences of values of x such that the corresponding sequences of values of $f(x)$ do not approach the same limit. In particular, if the sequence of values of x is $\{-1/n\}$, then the sequence of values of $f(x)$ in Fig. 3–2 is $\{1/n\}$ with limit zero. If the sequence of values of x is $\{1/n\}$, then the sequence of values of $f(x)$ is $\{n\}$, which increases without bound. Thus there exist null sequences of values of x such that the corresponding sequences of values of $f(x)$ do not have the same limit. The function and the curve in Fig. 3–2 are then said to be discontinuous at $x = 0$. In Fig. 3–1 every null sequence of values of x corresponds to a null sequence of values of $y = x \sin (1/x)$ and, assuming $y = 0$ when $x = 0$, the curve is said to be *continuous* at $x = 0$.

From Fig. 3–2 it is evident that if x approaches zero through any sequence of negative values, the corresponding sequence of values of $y = f(x)$ approaches zero. This common limit of all sequences for $f(x)$ corresponding to sequences of values of x which approach zero from the negative side is indicated by the notation $\lim_{x\to 0^-} f(x) = 0$. Similarly, from Fig. 3–2, we say that $\lim_{x\to 0^+} f(x)$ is unbounded, $\lim_{x\to 2^-} f(x) = \frac{1}{2}$, $\lim_{x\to 2^+} f(x) = \frac{1}{2}$, $\lim_{x\to -1^-} f(x) = 2$, $\lim_{x\to -1^+} f(x) = 1$. In the next section we shall use the concepts of limit from the left and limit from the right to define a continuous function.

Exercises

1. Sketch (algebraic conditions are not required) a function $f(x)$ such that
   (a) $\lim_{x\to 0^-} f(x) = -1$ and $\lim_{x\to 0^+} f(x) = +1$,
   (b) $\lim_{x\to 2^-} f(x) = 0$ and $\lim_{x\to 2^+} f(x) = 5$.
2. By definition, $\lim_{x\to a} f(x) = f(a)$ if and only if $\lim_{x\to a^-} f(x) = \lim_{x\to a^+} f(x) = f(a)$. Graph a function such that $\lim_{x\to a} f(x) = f(a)$ for every real value of a.
3. Graph a function such that $\lim_{x\to a} f(x) \neq f(a)$ when $a = -2, 0, 2$.
4. Graph a function such that $\lim_{x\to 0^-} f(x) = 0$, $f(0) = 1$, and $\lim_{x\to 0^+} f(x) = 2$.
5. Graph a function such that $\lim_{x\to 0^-} f(x) = 0$, $f(0) = 1$, and $\lim_{x\to 0^+} f(x) = 0$.

3–13 Continuous functions. When x is a continuous real variable, a single-valued function $y = f(x)$ defined on $a < x < b$ is said to be *continuous* at x_0, $a < x_0 < b$, if and only if $\lim_{x\to x_0^-} f(x) = f(x_0) = \lim_{x\to x_0^+} f(x)$. As mentioned previously, $f(x)$ is *continuous on an interval* if it is continuous at *every* point of the interval. Thus the function $f(x)$ is continuous on an interval if and only if at every point of that interval the limit from the left, the limit from the right, and the value of the function are all equal.

If the two limits $\lim_{x\to x_0^-} f(x)$ and $\lim_{x\to x_0^+} f(x)$ are finite and equal, then either their common limit is $f(x_0)$ and $f(x)$ is continuous at $x = x_0$, or their common limit is not $f(x_0)$ and $f(x)$ has a *removable discontinuity* at $x = x_0$. For example, the function

$$\begin{aligned} f(x) &= 1 && \text{when } x < 3, \\ &= 2 && \text{when } x = 3, \\ &= 4 - x && \text{when } x > 3 \end{aligned}$$

is graphed in Fig. 3–3. There is a removable discontinuity at $x = 3$ that may be removed by redefining $f(x)$ as

$$\begin{aligned} f(x) &= 1 && \text{when } x \leqq 3, \\ &= 4 - x && \text{when } x > 3. \end{aligned}$$

If the two limits are finite but are not equal, $f(x)$ has a discontinuity at $x = x_0$ that is called a *finite jump*. For example, in Fig. 3–2, $f(x)$ has a finite jump at $x = -1$.

If at least one of the limits is not finite, then either at least one limit becomes unbounded and $f(x)$ has an *infinite discontinuity* or at least one limit oscillates (does not exist) and $f(x)$ is discontinuous. For example, in Fig. 3–2, $f(x)$ has an infinite discontinuity at $x = 0$. The functions $y = 1/x$ and $y = 1/x^2$ also each have infinite discontinuities at $x = 0$. The function graphed in Fig. 3–4 is given by

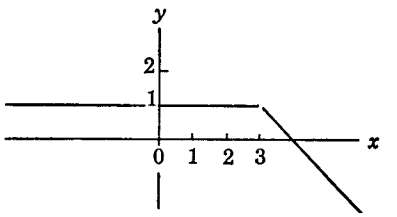


Fig. 3–3

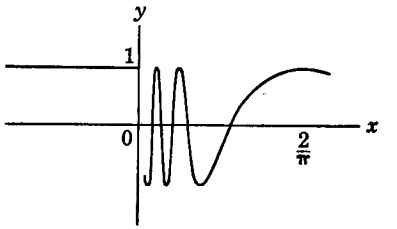


Fig. 3–4

$$\begin{aligned} f(x) &= 1 && \text{when } x \leqq 0, \\ &= \sin(1/x) && \text{when } x > 0. \end{aligned}$$

If $\{x\} = \{1/(n\pi)\}$, then $\{f(x)\} = \{0\}$ has limit 0; if

$$\{x\} = \{2/[(4n+1)\pi]\},$$

then $\{f(x)\} = \{1\}$ has limit 1. In fact, for any b, $-1 \leqq b \leqq 1$, there exists a null sequence of positive values of x such that the corresponding sequence of values of $f(x)$ has limit b. In this case, $\lim_{x\to 0^+} f(x)$ does not exist, and both the curve and the function are said to be discontinuous at $x = 0$.

We now give a second definition of a continuous function in which the concept of limit is replaced by some of the concepts used in the definition of a limit. Here again a figure is a very useful visual aid.

A single-valued function $f(x)$ (Fig. 3–5) defined for $a \leqq x \leqq b$ is continuous at $x = c$, $a \leqq c \leqq b$ if and only if for every $\epsilon > 0$ there exists a $\delta_{\epsilon c}$ such that $|f(x) - f(c)| < \epsilon$ for all x on $a \leqq x \leqq b$ satisfying $|x - c| < \delta_{\epsilon c}$. The notation $\delta_{\epsilon c}$ indicates that δ depends both upon the given ϵ and the chosen point $x = c$.

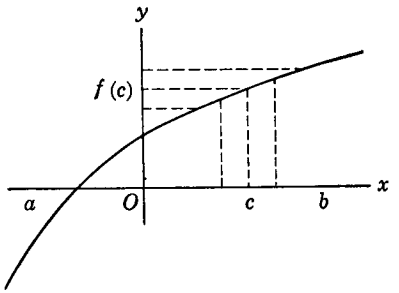


FIG. 3–5

If two functions $f(x)$ and $g(x)$ are each continuous at $x = c$, then by definition, given $\epsilon > 0$, there exists one positive $\delta_{\epsilon c}$ for $f(x)$ and another for $g(x)$. The smaller of these two positive numbers will be satisfactory for both $f(x)$ and $g(x)$ at $x = c$. Thus if $f(x)$ and $g(x)$ are both continuous, there exists a $\delta_{\epsilon c}$ such that both $|f(x) - f(c)| < \epsilon$ and $|g(x) - g(c)| < \epsilon$ for $|x - c| < \delta_{\epsilon c}$. We now use this fact to prove that the sum $f(x) + g(x)$ of two continuous functions is continuous. Given $\epsilon > 0$, we take $\epsilon/2$ as our positive number and choose δ such that for $|x - c| < \delta$, we have $|f(x) - f(c)| < \epsilon/2$ and $|g(x) - g(c)| < \epsilon/2$. Then

$$|[f(x) + g(x)] - [f(c) + g(c)]| = |f(x) - f(c) + g(x) - g(c)|$$
$$\leqq |f(x) - f(c)| + |g(x) - g(c)| < \epsilon/2 + \epsilon/2 = \epsilon.$$

This completes the proof that the sum of two continuous functions at a point is continuous at that point. The following sequence of exercises involves the above method of proof and leads to one of the principal results of this section, Exercise 4. This exercise can be proved from the preceding exercises, using the fact that each polynomial consists of a variable and a set of coefficients combined under ring operations.

EXERCISES

1. If $f(x)$ is continuous at $x = c$ and $g(x)$ is continuous at $x = c$, prove that $f(x) - g(x)$ is continuous at $x = c$.

2. If $f(x)$ and $g(x)$ are each continuous at $x = c$, prove that $f(x) \cdot g(x)$ is continuous at $x = c$.

3. If $f(x)$ and $g(x)$ are each continuous on a segment $a < x < b$, prove that $f(x) + g(x)$, $f(x) - g(x)$, and $f(x) \cdot g(x)$ are continuous on that segment.

4. Prove that any polynomial in a continuous real variable x with real coefficients is continuous for all real values of x.

5. If there exist real numbers a and b, $a < b$, such that $p(a) \cdot p(b) < 0$, where $p(x)$ is a real polynomial, prove that there is a real number c, $a < c < b$, such that $p(c) = 0$ [12; 66–67].

6. Graph and discuss continuity for each of the following:

(a) $y^2 = x^2 - 9$
(b) $y = \sqrt{x^2 - 9}$
(c) $y^2 = x^3$
(d) $x^2y = 1$
(e) $y = \dfrac{x^2 + 1}{x^2 - 1}$
(f) $y = 2^x$
(g) $y = 1$ when $x \leqq 0$, $= x$ when $x > 0$
(h) $y = |x|$ when $x \neq 0$, $= 1$ when $x = 0$

7. Draw graphs of single-valued functions illustrating each of the following:

(a) removable discontinuities at $x = 0$ and $x = 1$,
(b) finite discontinuities at $x = n$ for all positive integers n,
(c) infinite discontinuities at $x = +1$ and $x = -1$.

8. Give algebraic expressions for single-valued functions that could be used in Exercise 7.

9. A single-valued function $f(x)$ defined on the interval $a \leqq x \leqq b$ is *uniformly continuous* on that interval if and only if for every $\epsilon > 0$ there exists a δ_ϵ such that $|f(x) - f(x_0)| < \epsilon$ for all x and x_0 (any fixed value of x) on the given interval and satisfying $|x - x_0| < \delta_\epsilon$. Graph a function that is continuous on $0 < x < 1$ but not uniformly continuous on $0 \leqq x \leqq 1$.

10. Prove [12; 65–66] that if a function is continuous on a closed interval, it is uniformly continuous on that interval.

11. Prove that if a function is continuous on a closed interval, it (a) is bounded on that interval, (b) has a maximum M and a minimum m on that interval, and (c) takes on any value b, $m \leqq b \leqq M$, at least once on that interval.

12. Graph a function $y = f(x)$ that is single-valued, continuous, and increasing for $a \leqq x \leqq b$. It can be shown that there exists a corresponding inverse function $x = f^{-1}(y)$ that is also single-valued, continuous, and increasing [12; 67–68]. Check these properties for the function graphed.

13. Use the result stated in Exercise 12 to show that $y^{1/n}$ may be defined as a single-valued, continuous, and increasing function of y, where $y > 0$ and n is any positive integer (see Exercise 12, Section 3–10).

14. For $a > 1$ show that a^x is an increasing function of the positive real variable x (see Exercise 7, Section 3–11).

15. Show that $\log_a y$ may be defined as a single-valued, continuous, and increasing function of y, where $y > 0$ and $a > 1$.

3–14 Derivatives. The derivatives of a polynomial $p(x)$ may be defined in terms of limits or simply in terms of the coefficients of the polynomial $p(x + h)$. If $p(x) = x^2$, then $p(x + h) = x^2 + 2xh + h^2$.

The coefficient of h in $p(x+h)$ may be taken as the first derivative of $p(x) = x^2$. We write $(d/dx)p(x) = p'(x) = (x^2)' = 2x$. In general, the derivative of any polynomial $p(x)$ may be taken as the coefficient of h in the expansion of $p(x+h)$.

The *derivative* with respect to x of any polynomial $p(x)$ is usually defined as

$$p'(x) = \lim_{h \to 0} \frac{p(x+h) - p(x)}{h},$$

since for polynomials this limit exists for all real values of x and this definition may be readily extended to more general functions. The derivative at $x = x_0$ (a fixed value of x) of an arbitrary single-valued function $f(x)$ is defined by

$$f'(x_0) = \lim_{h \to 0} \frac{f(x_0+h) - f(x_0)}{h}$$

whenever the limit exists. When the limit does not exist, the derivative is undefined.

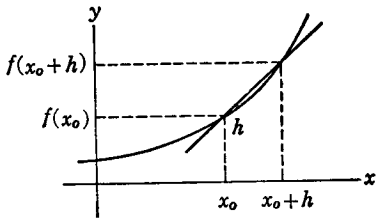


FIG. 3–6

Geometrically, the above definition of the derivative of $f(x)$ may be visualized in terms of tangent and secant lines on the graph of $f(x)$. Any line cutting the graph of $f(x)$ in two points may be called a secant line or a *secant*. In particular, we shall consider a secant (Fig. 3–6) cutting the graph of $f(x)$ at $[x_0, f(x_0)]$ and $[x_0+h, f(x_0+h)]$. The slope of this line may be taken as

$$\frac{f(x_0+h) - f(x_0)}{h}.$$

The tangent line or *tangent* to the graph of $f(x)$ at $[x_0, f(x_0)]$ may be defined as the limiting position of the above secant as h approaches zero. The slope of the tangent at $x = x_0$ is thus the limit as h approaches zero of the slope of the above secant and is precisely the derivative of $f(x)$ at $x = x_0$.

The derivative is undefined at every point of discontinuity and at points of a curve at which the slope of the tangent line is a discontinuous function of the independent variable. For example, if

$$\begin{aligned} f(x) &= x \qquad \text{when } x \leqq 1, \\ &= 2 - x \quad \text{when } x > 1, \end{aligned}$$

we have the graph in Fig. 3–7. At $x = 1$ the slope of the tangent changes abruptly from 1 to -1 and the derivative at $x = 1$ is undefined.

Given any monomial bx^n, we may obtain

$$b(x+h)^n = b[x^n + nhx^{n-1} + \frac{n(n-1)}{1 \cdot 2} h^2 x^{n-2} + \cdots + h^n].$$

Thus nbx^{n-1} is the derivative of bx^n whether the derivative is considered as a limit or as the coefficient of h. Next we observe that $[f(x)+g(x)]'$ may be taken as the coefficient of h in $[f(x+h)+g(x+h)]$, that is, $[f(x)+g(x)]' = f'(x)+g'(x)$. In other words, the derivative of the sum of two functions is equal to the sum of the derivatives of the functions. In particular, considering a polynomial as a sum of monomials, the derivative of a polynomial is equal to the sum of the derivatives of its terms (monomials). Since any monomial bx^n has derivative nbx^{n-1}, the derivative of any polynomial

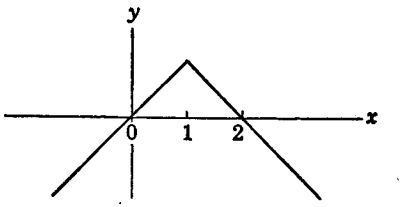


FIG. 3–7

$$p(x) = a_n + a_{n-1}x + a_{n-2}x^2 + a_{n-3}x^3 + \cdots + a_0 x^n$$

is given by

$$p'(x) = a_{n-1} + 2a_{n-2}x + 3a_{n-3}x^2 + \cdots + na_0 x^{n-1}.$$

We may thus find $p'(x)$ the first derivative with respect to x of any polynomial $p(x)$. Since $p'(x)$ is also a polynomial, the process may be repeated to obtain $p''(x)$, the derivative with respect to x of $p'(x)$,

$$p''(x) = 2a_{n-2} + 3 \cdot 2a_{n-3}x + \cdots + n(n-1)a_0 x^{n-2}.$$

Similarly, one may obtain $p'''(x)$, $p^{(4)}(x), \ldots, p^{(n)}(x)$. Finally, $p^{(n+k)}(x) = 0$ for any positive integer k, since the derivative of a constant is zero. We shall speak of $p^{(r)}(x)$, where r is any positive integer, as the rth derivative of $p(x)$.

The values of $p(x)$ and its first n derivatives when $x = 0$ are closely related to the coefficients of $p(x)$. For example, at $x = 0$, we have $p(0) = a_n$, $p'(0) = a_{n-1}$, $p''(0) = 2a_{n-2}$, $p'''(0) = 3(2a_{n-3})$, $\ldots$, $p^{(n)}(0) = n!(a_0)$. If these equations are solved respectively for a_n, a_{n-1}, $a_{n-2}, \ldots, a_0$, and the corresponding expressions substituted for the coefficients in the polynomial $p(x)$, we have

$$p(x) = p(0) + p'(0)x + \frac{p''(0)}{2!}x^2 + \frac{p'''(0)}{3!}x^3 + \cdots + \frac{p^{(n)}(0)}{n!}x^n.$$

This method of expressing a polynomial $p(x)$ in terms of its derivatives at a point (in the above case at $x = 0$) is a special case of *Taylor's formula* for polynomials:

$$p(x) = p(a) + p'(a)(x-a) + \frac{p''(a)}{2!}(x-a)^2 + \cdots + \frac{p^{(n)}(a)}{n!}(x-a)^n.$$

This formula may be used to express any polynomial $p(x)$ in terms of the values of the polynomial and its derivatives at $x = a$ for any real number a. It also gives an effective method for expressing $p(x)$ in the form $q(x+h)$, where $a = -h$, that is, replacing the variable x by the new variable $x + h$ (Section 3–8).

We have seen that given any polynomial $p(x)$ of degree n, we may obtain its kth derivative for any positive integer k. Also, for a polynomial $p(x)$ of degree n there are at most $n+1$ terms in its Taylor formula, since $p^{(n+k)}(x) = 0$ for all positive integers k. However, there exist functions such as $f(x) = e^x$ for which no derivative vanishes at $x = a$. In such cases Taylor's formula is extended to Taylor's series (Section 3–15).

We have now discussed all the properties of polynomials that we shall need in Chapter 4. The properties of the ring of polynomials corresponding to properties of the ring of integers, the fact that every polynomial in a continuous real variable is continuous, and the expression of any polynomial in terms of its derivatives in Taylor's formula will all be useful in our future discussions.

We shall conclude the present chapter with brief discussions of Taylor's series and analytic functions. Both concepts are very important in advanced mathematics but are not necessary for our future discussions. The fundamental role of analytic functions is evident from the correspondences between numbers and functions given at the end of Section 3–10.

Exercises

1. Derive Taylor's formula when $a \neq 0$.
2. Write $p(x) = x^3 - 5x^2 + 3x - 2$ as $q(x-1)$.
3. Write $p(x) = x^8 + 8x^7 - 6x^6 + 5x^2 + 3$ as $q(x+1)$.
4. Repeat Exercises 1, 2, and 3 of Section 3–8, using Taylor's formula.
5. Give a single-valued function of x that is continuous for all values of x but has no derivative at $x = 0$.

6. Prove that if $f(x)$ is an increasing function and $f'(x)$ exists on the interval $a < x < b$, then $f'(x) \geqq 0$ on $a < x < b$.
7. Graph an increasing function $f(x)$ on $a < x < b$, where $f'(d) = 0$ for some $x = d$, $a < d < b$.

***3–15 Taylor's series.** Given a polynomial $p(x)$ of degree m, we may assume (Section 3–8) that $p(x)$ may be expressed in the form

$$p(x) = b_0 + b_1(x-a) + \cdots + b_m(x-a)^m.$$

We may then determine the b's in terms of the value of $p(x)$ and its derivatives at $x = a$ as in Taylor's formula (Section 3–14). Only $(m+1)$ terms are necessary, since $p(x)$ has at most m derivatives different from zero.

Given any single-valued function $f(x)$, we may endeavor to express it in the form

$$f(x) = b_0 + b_1(x-a) + b_2(x-a)^2 + \cdots, \tag{3–4}$$

where there is a term associated with every non-negative, integral power of $(x-a)$. Such an expression is called an *infinite series* and corresponds in some respects to an infinite decimal in the development of our number system. If $f(x)$ may be expressed in the above form, the b's may again be expressed in terms of the values of $f(x)$ and its derivatives at $x = a$. For example, $b_0 = f(a)$, $b_1 = f'(a)$, $b_2 = f''(a)/2$. Thus a function must have derivatives of all orders if it is to be expanded as in (3–4). When the b's in (3–4) are replaced by the corresponding expressions in $f(x)$ and its derivatives at $x = a$, we have *Taylor's series:*

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)(x-a)^n}{n!} + \cdots.$$

In general, given any single-valued function $f(x)$ defined over an interval $c < x < d$, where $c < a < d$, we may consider in turn $f(a), f'(a), f''(a), \ldots, f^{(n)}(a), \ldots$. If $f^{(n)}(a)$ exists for all positive integral values of n, then $f(x)$ has a Taylor's series expansion at $x = a$, as indicated above.

The infinite series used in Section 1–16 may now be obtained. It is proved in all calculus textbooks that the derivative of e^x is e^x, the derivative of $\sin x$ is $\cos x$, and the derivative of $\cos x$ is $-\sin x$. Using $f(x) = e^x, f'(x) = e^x$, and Taylor's series at $x = 0$, we have $f^{(n)}(0) = 1$ for every positive integer n and

$$e^x = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \cdots.$$

When $f(x) = \sin x$, we have $f'(x) = \cos x, f''(x) = -\sin x$, $f'''(x) = -\cos x$, $f^{(4)}(x) = \sin x, \ldots$. At $x = 0$, $f(x) = f''(x) = f^{(2k)}(x) = 0$ and $f'(x) = 1, f'''(x) = -1, f^{(2k+1)}(x) = (-1)^k$ for any positive integer k. Then we have the Taylor series expansion:

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \cdots .$$

The Taylor series expansion of $\cos x$ may be similarly obtained (Exercise 1).

We have formally indicated how to obtain a Taylor series expansion at $x = a$ of any function that has derivatives of all orders at $x = a$. The question of convergence of the infinite series, i.e., for what values of x the series has significance, will be left for textbooks in analysis such as [12; 320–329, 365–424]. The series for e^x and $\sin x$ converge for all real values of x. Our interest in the formal expansion will be evident from the definition of an analytic function.

Exercises

1. Develop a Taylor series expansion of $\cos x$ in terms of powers of x.
2. Use the Taylor series expansions of $\cos x$ and $\sin x$ to form a Taylor series expansion of $\cos x + i \sin x$.
3. Develop a Taylor series expansion of e^{ix} and compare with that found in Exercise 2.

***3–16 Analytic functions.** We now complete our development of the correspondences between numbers and functions listed in Section 3–10. We defined a polynomial in Section 3–1 and have noted the correspondences between polynomials and integers in Sections 3–1 to 3–9. The correspondence between rational numbers and rational functions was established in Section 3–3. The remaining correspondences will now be briefly discussed.

A number is said to be algebraic if it satisfies a polynomial equation with integral coefficients that is not identically zero (Section 1–10). All other numbers are said to be transcendental. Similarly, a function $y = f(x)$ is said to be an *algebraic function* of x if it satisfies a polynomial equation with polynomials in x as coefficients that is not identically zero in y. For example, $y = \sqrt[3]{x}$ satisfies $y^3 - x = 0$ and is therefore an algebraic function. All functions that are not algebraic are called *transcendental functions*. For example, $\sin x$, $\log x$, and e^x are transcendental functions.

We obtained real numbers (Section 1–10) by assuming that all decimals exist as numbers. The set of real numbers could also have been obtained by assuming that every Cauchy sequence of rational numbers represents a real number. We now obtain analytic functions of a variable x by assuming that every Taylor series in x represents a function. Many texts define an *analytic function* of a variable x to be a function that has a Taylor series expansion. Here, as in Section 3–15, there is the problem of the convergence of the series. In the above statements it is assumed that a Taylor series expansion exists if and only if it has meaning on some segment $a < x < b$.

We have now seen that polynomials form not only a ring with properties very similar to those of the ring of integers, but also that the set of polynomials may be extended to the set of analytic functions in a manner very similar to that used when the integers were extended to the set of real numbers. We shall consider additional properties of polynomials in our study of the theory of polynomial equations in Chapter 4.

Exercises

1. List ten rational functions of x and state the values of x for which each is defined (Section 3–3).
2. List ten algebraic functions and give a polynomial equation that each satisfies.
3. List ten transcendental functions.

CHAPTER 4

THEORY OF EQUATIONS

We have previously discussed the operations, monomials, and relations used in the formation of polynomials. In this chapter we shall consider equations $p(x) = 0$ obtained by setting a polynomial in one variable equal to zero. For example, the polynomial equation $x^2 - x - 6 = 0$ is satisfied when $x = 3$ or $x = -2$. The above equation thus places a condition upon the variable and is called a *conditional equation*. The equation $(x - 1)^2 = x^2 - 2x + 1$ is an identity (Section 3–4) and is satisfied by all numerical values of the variable x. We shall be primarily concerned with methods of determining the numbers that, when substituted for x in the equation $p(x) = 0$, will make the equality hold. Such numbers are called *zeros* of the polynomial or *roots* of the equation. They are also sometimes called *solutions* of the equation. The topics considered in this chapter are selected to prepare the reader

(i) to find all integral and rational solutions of any given polynomial equation in one variable with rational coefficients,

(ii) to find all solutions of any given polynomial equation with complex coefficients whenever possible using radicals and the four rational operations,

(iii) to determine on any given interval the exact number of real solutions of any given polynomial equation with real coefficients, and

(iv) to approximate as closely as desired any real solution of a given polynomial equation with real coefficients.

4–1 Zeros of a polynomial. A number b has been defined above to be a zero of the polynomial $p(x)$ if and only if $p(b) = 0$. Given any polynomial $p(x)$, we may seek numbers b such that $p(b) = 0$, that is, seek the zeros of the polynomial. Given any polynomial $p(x)$ and a number b, we may seek necessary and sufficient conditions that $p(b) = 0$. Of course, one such condition would be obtained by outright substitution of b for x in the polynomial to obtain the number $p(b)$ and observe whether or not that number is zero. However, there exists another necessary and sufficient condition (Theorem 4–1)

that is often much easier to apply (Section 4–2) than straight substitution.

Given a polynomial $p(x)$ and a number b, we may find by long division a polynomial $q(x)$ and a constant R such that

$$p(x) = (x - b) \cdot q(x) + R, \tag{4–1}$$

as prescribed by the Division Algorithm (Section 3–5) for polynomials in one variable. The identity (4–1) gives $p(b) = R$ when $x = b$. This fact is sometimes referred to as the

REMAINDER THEOREM: *The remainder when $p(x)$ is divided by $x - b$ is $p(b)$.*

We may also obtain the following theorem from (4–1).

THEOREM 4–1. FACTOR THEOREM: *A polynomial $p(x)$ vanishes for $x = b$ if and only if it is divisible by $x - b$.*

In other words, the polynomial equation $p(x) = 0$ has a root $x = b$ if and only if the polynomial $p(x)$ has a divisor or factor $x - b$, that is, if and only if $R = 0$ in (4–1). The task of finding R and $q(x)$ in (4–1) can often be most easily and quickly performed by synthetic division.

EXERCISES

1. Give three examples of the Remainder Theorem where $p(x)$ has degree at least three.
2. Repeat Exercise 1 for the Factor Theorem.
3. Give four conditional equations.
4. Give three equations that are identities.

4–2 Synthetic division. Synthetic division, like the array used to find the greatest common divisor of two integers (Section 2–5), gives an elementary and concise method of presenting the results of algebraic calculations.

Suppose we are given a number b and a polynomial of degree $n > 0$, say

$$p(x) = b_0x^n + b_1x^{n-1} + \cdots + b_n. \tag{4–2}$$

The polynomial $q(x)$ in (4–1) must then have degree $n - 1$ and thus be of the form

$$q(x) = c_0x^{n-1} + c_1x^{n-2} + \cdots + c_{n-1},$$

where the c's are to be determined. The identity (4–1) then has the form

$$(4\text{–}3) \qquad b_0x^n + b_1x^{n-1} + \cdots + b_n = c_0x^n + (c_1 - c_0b)x^{n-1} + \cdots + (c_{n-1} - c_{n-2}b)x + R - c_{n-1}b.$$

When $x = 0$, (4–3) has the form $b_n = R - c_{n-1}b$. If these equal constant terms are subtracted from both sides of (4–3) and both sides of the resulting equation are divided by x, we find, by again putting $x = 0$, that $b_{n-1} = c_{n-1} - c_{n-2}b$. In general, coefficients of like powers of x must be equal, and we have

$$\begin{aligned} b_0 &= c_0, \\ b_1 &= c_1 - c_0b, \\ b_2 &= c_2 - c_1b, \\ &\cdot \\ &\cdot \\ &\cdot \\ b_{n-1} &= c_{n-1} - c_{n-2}b \\ b_n &= R - c_{n-1}b. \end{aligned}$$

These equations may be solved for the c's to give

$$\begin{aligned} c_0 &= b_0, \\ c_1 &= b_1 + c_0b, \\ c_2 &= b_2 + c_1b, \\ &\cdot \\ &\cdot \\ &\cdot \\ c_{n-1} &= b_{n-1} + c_{n-2}b, \\ R &= b_n + c_{n-1}b, \end{aligned}$$

and thus successively determine the coefficients of $q(x)$ and R.

The above relations may be very concisely expressed using the following array, i.e., by *synthetic division*,

$$\begin{array}{cccccc|c} b_0 & b_1 & b_2 & \cdots & b_{n-1} & b_n & b \\ 0 & c_0b & c_1b & \cdots & c_{n-2}b & c_{n-1}b & \\ \hline c_0 & c_1 & c_2 & \cdots & c_{n-1} & R & \end{array}$$

In this array the first row contains the coefficients of $p(x)$ (including all zero coefficients); the first element of the second row is zero and each succeeding element is the product of the number b and the element of the third row in the immediately preceding column; each element of the third row is the sum of those directly above it. For

example, if $p(x) = x^3 - 3x^2 + 7x - 10$ and $b = 4$, the above array becomes

$$\begin{array}{cccc|c} 1 & -3 & 7 & -10 & 4 \\ 0 & 4 & 4 & 44 & \\ \hline 1 & 1 & 11 & 34 & \end{array}$$

from which we conclude that $q(x) = x^2 + x + 11$ and $R = 34$, that is, $x^3 - 3x^2 + 7x - 10 = (x - 4)(x^2 + x + 11) + 34$. Since the first element, zero, of the second row is always the same, it is frequently not written down.

The above array for the division of a polynomial $p(x)$ by a monic linear polynomial $x - b$ may be modified to include the division of $p(x)$ by a quadratic or higher degree polynomial in x [47; 56–58]. In general, synthetic division is very useful in checking that $p(b) = 0$ as above; expressing $p(x)$ in the form $q(x - b)$, that is, reducing the values of the zeros by b (Sections 3–8 and 4–3); solving cubic and quartic equations (Sections 4–9 and 4–10); computing a table of values for graphing $y = p(x)$; determining an upper bound for the zeros of $p(x)$ (Section 4–11); and solving numerical equations (Section 4–13).

Exercises

1. Without actual division, find the remainder when
   (a) $x^2 - 5x + 6$ is divided by $x - 4$,
   (b) $x^3 - 3x^2 + 6x - 5$ is divided by $x - 3$,
   (c) $x^4 - 3x^2 - 2x - 4$ is divided by $x + 3$.
2. Without actual division, show that
   (a) $13x^{10} + 14x^5 + 1$ is divisible by $x + 1$,
   (b) $2x^4 - x^3 - 6x^2 + 4x - 8$ is divisible by $x - 2$ and by $x + 2$,
   (c) $v^4 - 3v^3 + 3v^2 - 3v + 2$ is divisible by $v - 1$ and by $v - 2$,
   (d) $x^n - 1$ is divisible by $x - 1$.
3. By synthetic division, find the quotient and remainder when
   (a) $2x^4 + 4x^3 - x^2 - 16x - 12$ is divided by $x + 4$,
   (b) $3x^4 - 27x^2 + 14x + 120$ is divided by $x - 6$,
   (c) $x^4 - 4x^3 - 8x + 32$ is divided by $x - 4$.
4. Given that $p(x) = x^3 - x^2 - 4x - 6$ has a zero at $x = 3$, find a quadratic polynomial having as zeros the other two zeros of $p(x)$.
5. Use synthetic division to write each of the following in the form (4–1):
   (a) $p(x) = x^3 - 3x^2 + 2x + 1$, $b = 1$,
   (b) $p(x) = x^6 - x^3 + 2x^2 - 77$, $b = -1$,
   (c) $p(x) = x^5 - 7x^4 + 3x^3 - 5x^2 + 6$, $b = 5$,
   (d) $p(y) = y^5 + 4y^3 + 2$, $b = -3$,
   (e) $p(y) = y^6 + 1$, $b = -2$.

4–3 Change of variable. In Section 3–8 we proved that any polynomial $p(x)$ could be expressed in the form $q(x-b)$. In this section we shall see how synthetic division may be used to find the polynomial $q(x-b)$ when the polynomial $p(x)$ and a number b are given. One method, using Taylor's formula, has already been discussed in Section 3–14.

Given a polynomial $p(x)$ of degree m and a number b, we may (Theorem 3–4) express $p(x)$ in the form

$$p(x) = a_0 + a_1(x-b) + a_2(x-b)^2 + \cdots + a_m(x-b)^m, \tag{4–4}$$

where the a's are to be determined. Then

$$\begin{aligned} p(x) &= (x-b)[a_1 + a_2(x-b) + \cdots + a_m(x-b)^{m-1}] + a_0 \\ &= (x-b)\cdot q_1(x) + a_0, \end{aligned}$$

as in (4–1). Thus $a_0 = p(b)$ may be computed by synthetic division, as in Section 4–2. Next, we write

$$\begin{aligned} q_1(x) &= (x-b)[a_2 + a_3(x-b) + \cdots + a_m(x-b)^{m-2}] + a_1 \\ &= (x-b)\cdot q_2(x) + a_1, \end{aligned}$$

whence $a_1 = q_1(b)$. This process may be continued to give $a_2 = q_2(b)$, $\ldots$, $a_m = q_m(b)$ and thus completely determine the a's in (4–4), i.e., completely determine $q(x-b) = p(x)$.

Suppose that $p(x) = x^4 - x^3 - 4x^2 + 3x - 1$ and $b = 2$. We may use the first three rows of the following array to find

$$q_1(x) = x^3 + x^2 - 2x - 1$$

and $a_0 = -3$. Then we may continue the same array for two more rows to find $q_2(x) = x^2 + 3x + 4$ and $a_1 = 7$. Similarly, $q_3(x) = x + 5$, $a_2 = 14$; $q_4(x) = 1$, $a_3 = 7$; $a_4 = 1$.

					⌊ 2
1	−1	−4	3	−1	
0	2	2	−4	−2	
1	1	−2	−1	−3	
0	2	6	8		
1	3	4	7		
0	2	10			
1	5	14			
0	2				
1	7				
0					
1					

In general, we may write any polynomial $p(x)$ of degree m as $q(x-b)$ and use synthetic division $(m+1)$ times to find the coefficients of the new polynomial $q(x-b)$. Thus the theory of Section 3–8 can now be carried out for any given polynomial $p(x)$ and any given number b. Note that this reduction of the zeros or change of variable is accomplished without any reference to the values of the zeros of the polynomial.

Exercises

1. Do Exercise 1 of Section 3–8, using the above method. Compare this method with those used in Sections 3–8 and 3–14.
2. Do Exercises 2 and 3 of Section 3–8, using the above method.
3. Write $x^6 - 7x^5 + x^4 - 3x^2 + 11$ as $q(x-2)$.
4. Write $x^7 - 1$ as $q(x-1)$.
5. Write $x^7 - 1$ as $q(x+1)$.
6. Find equations whose roots are
   (a) two less than those of $x^3 - 7x^2 + 2x + 1 = 0$.
   (b) one more than those of $x^4 + 3x^3 - 5x^2 + x + 7 = 0$.

4–4 Number of roots. The most practical aspect of this chapter is that concerned with finding one or more roots of a polynomial equation. In general, we say that a polynomial equation is *solved* when all its roots have been determined. Thus before solving a polynomial equation it is desirable to know the total number of roots that are required. This number can be stated in advance for any given polynomial with complex coefficients. If the polynomial is a constant b, the equation has no roots if $b \neq 0$, and it has every complex number as a root if $b = 0$. If the polynomial is not a constant, we define the degree of the polynomial equation $p(x) = 0$ to be the same as the degree of $p(x)$ and have

Theorem 4–2. *Every polynomial equation of degree $m > 0$ with complex coefficients has precisely m complex roots (not necessarily distinct).*

This theorem may be easily proved in the theory of functions of a complex variable. We shall not attempt to give a complete proof here since the algebraic proofs are all somewhat involved. Instead we shall assume the following theorem and use it to prove Theorem 4–2.

Theorem 4–3. Fundamental Theorem of Algebra. *Every polynomial $p(x)$ of positive degree with complex coefficients has at least one complex zero.*

Theorems 4–2 and 4–3 are very closely related to the fact that the set of complex numbers is algebraically closed, as mentioned in the optional Section 1–18 and assigned as Exercise 6 of that section. For present purposes any reader who omitted that exercise should consult another text or accept Theorem 4–3 without formal proof.

Now we use Theorem 4–3 and prove Theorem 4–2. Given any polynomial $p(x)$ of degree m, we may, by Theorem 4–3, designate one of its zeros by the complex number r_1 and, by Theorem 4–1, write $p(x) = (x - r_1)p_1(x)$, where $p_1(x)$ is a polynomial of degree $m - 1$. The coefficients of $p_1(x)$ may be found by synthetic division and expressed in terms of the zero r_1, the coefficients of $p(x)$, and the three ring operations (Section 4–2). Therefore, the coefficients of $p_1(x)$ are complex numbers, and the above procedure may be repeated for $p_1(x)$ if $m - 1 > 0$. Since the degree m of $f(x)$ is finite, this procedure may be repeated only a finite number of times, giving

$$\begin{aligned} p(x) &= (x - r_1)p_1(x), \\ p_1(x) &= (x - r_2)p_2(x), \\ p_2(x) &= (x - r_3)p_3(x), \\ &\cdot \\ &\cdot \\ &\cdot \\ p_{n-1}(x) &= (x - r_m)a_0, \end{aligned}$$

and

$$p(x) = a_0(x - r_1)(x - r_2)(x - r_3) \dots (x - r_m), \tag{4–5}$$

where a_0 is the initial of $p(x)$.

The Unique Factorization Theorem states for polynomials (Exercise 9, Section 3–6) that every factor of $p(x)$ of the form $x - b$ is included in (4–5). Theorem 4–1 then states that the zeros of $p(x)$ are precisely $r_1, r_2, r_3, \dots, r_m$. This completes the proof of Theorem 4–2 using Theorem 4–3. For the remainder of this chapter we shall be primarily concerned with the determination of the roots of polynomial equations.

Exercises

Prove the following statements.

1. Any polynomial equation of the form $d_0x^m + d_1x^{m-1} + \cdots + d_m = 0$ having more than m distinct roots is identically zero, that is, $d_i = 0$ for $i = 0, 1, 2, \dots, m$.

2. If two polynomials $p(x)$ and $q(x)$ of degree m are equal for more than m distinct values of x, they are identical.

3. If two polynomial equations of degree m have precisely the same roots, they are associates (Section 3–4). [*Hint:* Consider $f(x) + kg(x)$, where k is chosen so that the terms of degree m drop out.]

4–5 Determination of the roots. We have seen that every polynomial equation of degree m with complex coefficients has exactly m complex roots (not necessarily distinct). There still remains the very practical problem of finding the roots of a given polynomial equation $p(x) = 0$.

The general linear equation is of the form $ax + b = 0$, where $a \neq 0$. It has a unique root $x = -b/a$. Every linear equation with rational coefficients has a rational root; every linear equation with complex coefficients (real or imaginary) has a complex root.

The general quadratic equation is of the form $ax^2 + bx + c = 0$, where $a \neq 0$. Its two roots may be found by completing the square of the left member as follows:

$$\begin{aligned} x^2 + \frac{b}{a}x &= -\frac{c}{a}, \\ x^2 + \frac{b}{a}x + \frac{b^2}{4a^2} &= \frac{b^2}{4a^2} - \frac{c}{a}, \\ x + \frac{b}{2a} &= \pm\frac{\sqrt{b^2 - 4ac}}{2a}, \\ x &= \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}. \end{aligned}$$

The two roots of $ax^2 + bx + c = 0$, where $a \neq 0$, are then

$$(-b + \sqrt{b^2 - 4ac})/2a \text{ and } (-b - \sqrt{b^2 - 4ac})/2a.$$

The number $b^2 - 4ac$ is called the *discriminant* of the quadratic equation. If the two roots are respectively designated by r_1 and r_2, it is easily verified that they satisfy the elementary symmetric polynomial relations (Section 4–7) $r_1 + r_2 = -b/a$, $r_1r_2 = c/a$. If the coefficients are real numbers, we have

THEOREM 4–4. *A quadratic equation with real coefficients has two roots that are real and unequal, real and equal, or conjugate imaginary according as the discriminant is positive, zero, or negative.*

The roots of the general cubic (Section 4–9) and quartic (Section 4–10) equations may be expressed in terms of the coefficients, using

the four fundamental operations and radicals. The general equation of degree m, $m > 4$, cannot be solved in terms of the coefficients, using the four fundamental operations, and radicals. This negative result can be proved using the work of Evariste Galois in the theory of groups. Lillian Lieber has written a very interesting and readable booklet [32] giving this result and some other applications of Galois' theories.

The approximate value or graphical location of the real zeros of a polynomial $p(x)$ can often be determined from the continuity of $p(x)$. By Exercise 4, Section 3–13, every polynomial in the continuous real variable x is continuous for all finite values of x. Thus (Exercise 5, Section 3–13) if there exist real numbers a and b, $a < b$, such that $p(a) \cdot p(b) < 0$, there is a real number c, $a < c < b$, such that $p(c) = 0$. Also, since $p(x)$ has the sign of its initial (Section 3–1) for sufficiently large positive values of x, every real polynomial equation of odd degree and positive initial has at least one real root opposite in sign to its last term; every real polynomial equation of even degree whose initial and constant term have opposite signs has at least one positive root and at least one negative root.

We shall consider other methods for approximating the roots of a polynomial equation in Section 4–14. First we consider further some of the relations between the coefficients and the roots.

Exercises

Describe the roots of the equations in Exercises 1–7.

1. $5x - 17 = 0$
2. $ix + 7i + 5 = 0$
3. $x^2 - 3x + 7 = 0$
4. $2x^2 - 7x + 35 = 0$
5. $x^3 - 3x^2 + 3x - 1 = (x - 1)^3$
6. $5x^2 - 3x - 2 = 0$
7. $x^2 + 2x + 7 = x(x + 2)$.

8. Construct a quadratic equation with roots whose sum is 3 and whose product is 5. Is there a unique answer?

9. Find the sum and product of the roots in each of the following equations:

   (a) $x^2 - 5x + 6 = 0$
   (b) $2x^2 - 3x + 5 = 0$
   (c) $3x^2 + 4x - 7 = 0$

10. Describe (without finding them) the roots in each part of Exercise 9.

4–6 Conjugate imaginary roots. We shall prove the following theorem:

THEOREM 4–5. *The imaginary roots of a polynomial equation with real coefficients occur in conjugate pairs.*

Let $p(z)$ be a polynomial with real coefficients and suppose that $p(w) = 0$, where $w = a + bi$; a, b are real; $b \neq 0$. We shall prove that $p(\overline{w}) = 0$ where $\overline{w} = a - bi$. The quadratic polynomial

$$(z - w)(z - \overline{w}) = z^2 - 2az + a^2 + b^2$$

may be used with $p(z)$ in the Division Algorithm (Section 3–5) to obtain

$$p(z) = [z^2 - 2az + a^2 + b^2] \cdot g(z) + sz + t,$$

where $g(z)$ is a polynomial. At $z = w$ we have $p(w) = 0 = 0 + sw + t$ or, using $w = a + bi$, $0 = sa + sbi + t$. Equating real and imaginary parts of this equation gives $sa + t = 0$ and $sb = 0$, whence $s = 0$ and $t = 0$, since $b \neq 0$. Then $p(z) = (z - w)(z - \overline{w}) \cdot g(z)$, $p(\overline{w}) = 0$, and the proof of Theorem 4–5 is complete.

In Section 4–4 we found that every polynomial $p(x)$ of degree m with complex coefficients had m complex roots. This implied that $p(x)$ could be written as a product of m linear factors with complex coefficients (4–5). In other words, every irreducible polynomial whose factors may have arbitrary complex coefficients is linear. Theorem 4–5 and the fact that $(z - w)(z - \overline{w})$ is a quadratic polynomial with real coefficients now imply that every irreducible polynomial whose factors may have arbitrary real coefficients is quadratic or linear. A polynomial of any degree m may be irreducible when the factors are required to have integral or rational coefficients. For example, $x^m - 2$ for any positive integer m has no factors of degree less than m with rational coefficients.

The solution of a polynomial equation and the factorization of the polynomial into irreducible factors are therefore equivalent, in the sense of Theorem 4–1, only when the coefficients of the factors may be arbitrary complex numbers. Actually, the set of algebraic complex numbers is sufficient, as mentioned in Section 1–18.

Exercises

1. Prove that the quadratic surd (irrational) roots $a + \sqrt{b}$ of a polynomial equation with rational coefficients occur in conjugate pairs.

2. Form a rational cubic equation having 1 and $3 - 2\sqrt{-1}$ as two of its roots.

3. Form a real quartic equation having $1 + 5\sqrt{-1}$ and $5 - \sqrt{-1}$ as two of its roots.

4. Given that $x^3 - 3x^2 - 2x + 6 = 0$ has a root $x = \sqrt{2}$, find the other two roots.

5. Given that $2z^4 - 12z^3 + 19z^2 - 6z + 9 = 0$ has $i/\sqrt{2}$ as a root, find the other three roots.

4–7 Elementary symmetric polynomials. We have seen that any quadratic equation $ax^2 + bx + c = 0$ has two roots r, s which satisfy the relations $r + s = -b/a$, $rs = c/a$ (Section 4–5). In general, if $p(x)$ is a polynomial of degree m with complex coefficients, the polynomials in the zeros of $p(x)$ obtained by taking the sum of all the zeros of $p(x)$, the sum of all products of pairs of zeros of $p(x)$, the sum of all products of triples of zeros of $p(x)$, . . . , the product of all zeros of $p(x)$ may be expressed rationally in terms of the coefficients of $p(x)$. These polynomials in the zeros of a polynomial $p(x)$ are called *elementary symmetric polynomials.*

Since any m numbers may be considered as the zeros of a polynomial of degree m, we may consider the elementary symmetric polynomials of any m given numbers. For example, the numbers 1, 2, 3 are zeros of the polynomial $p(x) = (x - 1)(x - 2)(x - 3) = x^3 - 6x^2 + 11x - 6$. The elementary symmetric polynomials of these numbers or of the zeros of $p(x)$ may be expressed in terms of the coefficients of $p(x)$ as follows: $1 + 2 + 3 = 6$, the coefficient of x^2 with its sign changed; $1 \cdot 2 + 1 \cdot 3 + 2 \cdot 3 = 11$, the coefficient of x; and $1 \cdot 2 \cdot 3 = 6$, the constant term of $p(x)$ with its sign changed. In general, when the set of allowable coefficients forms a field, we may divide any polynomial $p(x)$ of degree m by its initial to obtain a monic polynomial having the same zeros as $p(x)$. In the monic polynomial the sum of the m zeros is equal to the coefficient of x^{m-1} with its sign changed; the sum of the products of the zeros in pairs is equal to the coefficient of x^{m-2}; . . . ; the product of the zeros is equal to the constant term multiplied by $(-1)^m$.

These general results may be obtained using the expansion (4–5) of $p(x)$:

$$p(x) = a_0(x - r_1)(x - r_2)(x - r_3) \cdots (x - r_m).$$

Then

$$\begin{aligned} p(x) &= a_0[x^m - (r_1 + r_2 + \cdots + r_m)x^{m-1} \\ &\qquad + (r_1r_2 + r_1r_3 + r_2r_3 + \cdots + r_{m-1}r_m)x^{m-2} - \cdots \\ &\qquad + (-1)^m r_1r_2 \ldots r_m] \\ &= a_0[x^m - S_1x^{m-1} + S_2x^{m-2} - S_3x^{m-3} + \cdots + (-1)^m S_m], \end{aligned}$$

where S_j is the elementary symmetric polynomial of degree j, that is,

$$\begin{aligned} S_1 &= r_1 + r_2 + \cdots + r_m = \Sigma r_i, \\ S_2 &= r_1r_2 + r_1r_3 + r_2r_3 + r_1r_4 + \cdots + r_{m-1}r_m = \Sigma r_ir_j, \\ &\cdot \\ &\cdot \\ &\cdot \\ S_j &= \Sigma r_{i1}r_{i2} \ldots r_{ij}, \\ &\cdot \\ &\cdot \\ &\cdot \\ S_m &= r_1r_2r_3 \ldots r_m. \end{aligned}$$

Given any polynomial of degree m, we may obtain a corresponding monic polynomial in which the sum of the zeros is equal to the coefficient of x^{m-1} with its sign changed; the product of the zeros is $(-1)^m$ times the constant term; and, in general, S_k is $(-1)^k$ times the coefficient of x^{m-k}. For example, if $p(x) = (x + 1)(x - 1)(x + 2)(x - 2)$ with roots -1, 1, -2, 2, then $p(x) = x^4 - 5x^2 + 4$, where the leading coefficient is unity, the coefficient of x^3 is $0 = -[(-1) + 1 + (-2) + 2]$, the coefficient of x^2 is

$$-5 = (-1) \cdot 1 + (-1)(-2) + (-1) \cdot 2 + 1 \cdot (-2) + 1 \cdot 2 + (-2) \cdot 2,$$

the coefficient of x is

$$0 = -[(-1) \cdot 1 \cdot (-2) + (-1) \cdot 1 \cdot 2 + (-1)(-2) \cdot 2 + 1 \cdot (-2) \cdot 2],$$

and the constant term is $4 = (-1) \cdot 1 \cdot (-2) \cdot 2$.

The elementary symmetric polynomials S_j may be used to solve polynomial equations when some additional relation is known among the roots. For example, if it is known that two of the roots of the equation $x^3 - 4x^2 + 5x - 2 = 0$ are equal, then the three roots may be represented by r, r, and s. The elementary symmetric polynomials may now be used to complete the solution of the equation. Using S_1 and S_2, we have $2r + s = 4$ and $2rs + r^2 = 5$. These two equations may be solved simultaneously by substituting from the linear into the quadratic to obtain $2r(4 - 2r) + r^2 = 5$, $3r^2 - 8r + 5 = 0$, whence $r = 1$ and $s = 2$ or $r = \frac{5}{3}$ and $s = \frac{2}{3}$. The second pair of values does not give the correct value for S_3, since we must have $r^2s = 2$, and thus the given equation has roots 1, 1, and 2.

When nothing is known about the roots of a polynomial equation except that they are roots of $p(x) = 0$, the use of the elementary symmetric polynomials merely leads to another equation that is

essentially equivalent to $p(x) = 0$. For example, suppose that $x^2 + bx + c = 0$ has roots r and s. Then $r + s = -b$, $rs = c$, and, by substitution, $r(-b - r) = c$, or $r^2 + br + c = 0$.

The following exercises include several applications of the above results. We shall consider other important uses of the elementary symmetric polynomials in the next section.

Exercises

1. Without using linear factors, find
   (a) a cubic equation having roots 1, 2, 3,
   (b) a cubic equation having roots 0, -2, 2,
   (c) a quartic equation having roots 2, 2, -2, -2.
2. Given that $x^4 + 14x^3 + 73x^2 + 168x + 144 = 0$ has two pairs of double roots, find its roots.
3. Given that one root of $x^3 - 27x^2 + 242x - 720 = 0$ is equal to half the sum of the other two, find the roots.
4. Given that two of the roots of $x^3 + 7x^2 - 6x - 72 = 0$ are in the ratio of 3 to 2, find all three roots.
5. Given that one root is the negative of another, solve
   (a) $4x^3 - 12x^2 - 25x + 75 = 0$,
   (b) $4x^3 - 16x^2 - 9x + 36 = 0$.
6. What relations must exist between the coefficients of a general quadratic equation if one root is twice the other?
7. Solve $x^3 + 7x^2 - 21x - 27 = 0$, given that the roots are in a geometric progression.
8. Solve $x^3 - 3x^2 - 13x + 15 = 0$, given that the roots are in an arithmetic progression.

4–8 Transformations of roots. This section is essentially an extension of Sections 3–8 and 4–3. In Section 3–8 we found that any polynomial $p(x)$ could theoretically be expressed in the form $q(ax + b)$, where $a \neq 0$. In Section 4–3 we used synthetic division to obtain the new polynomial in the special case $a = 1$. In this section we shall use the elementary symmetric polynomials to obtain the new polynomial $q(ax + b)$ for any $a \neq 0$, i.e., given a polynomial $p(x)$ with zeros r_j $(j = 1, 2, \ldots, m)$ we shall find a polynomial $q(y)$ with zeros $ar_j + b$ for any given numbers $a \neq 0$ and b.

Any polynomial with coefficients from a field and zeros $r_1, r_2, \ldots, r_m$ has an associate of the form

$$p(x) = x^m - S_1x^{m-1} + S_2x^{m-2} + \cdots + (-1)^jS_jx^{m-j} + \cdots + (-1)^mS_m,$$

where the S_j are the elementary symmetric polynomials of degree j (Section 4–7). If we multiply each r_j by a number k, then

$$kr_1 + kr_2 + \cdots + kr_m = k(r_1 + r_2 + \cdots + r_m) = kS_1,$$
$$(kr_1)(kr_2) + (kr_1)(kr_3) + (kr_2)(kr_3) + \cdots + (kr_{m-1})(kr_m) = k^2S_2,$$
$$\cdots$$
$$(kr_1)(kr_2)(kr_3) \ldots (kr_m) = k^mS_m.$$

Thus multiplication of the zeros of $p(x)$ by k results in the multiplication of the elementary symmetric polynomial S_j by k^j. Conversely, if we multiply the S_j by k^j, we obtain a new polynomial having zeros exactly k times those of $p(x)$. For example, $x^2 - 4x + 3$ has zeros 1 and 3, where $S_1 = 4$ and $S_2 = 3$. If we construct a new polynomial, using $2S_1 = 8$ and $2^2S_2 = 12$ as the new elementary symmetric polynomials, we obtain $q(y) = y^2 - 8y + 12$ with zeros 2 and 6. In general, we have

THEOREM 4–6. *Given a polynomial $p(x) = a_0x^m + a_1x^{m-1} + \cdots + a_m$ and a number k, we may immediately write down a polynomial $q(y) = a_0y^m + a_1ky^{m-1} + a_2k^2y^{m-2} + \cdots + a_mk^m$ with zeros k times those of $p(x)$.*

In the following theorem, Theorem 4–7, we shall discuss a procedure for obtaining a polynomial with zeros that are the reciprocals of the zeros of a given polynomial. For example, given the polynomial $x^2 - x - 6$ with zeros 3 and -2, we may, by Theorem 4–7, obtain $1 - x - 6x^2$ with zeros $\frac{1}{3}$ and $-\frac{1}{2}$. In connection with this theorem it is necessary to consider that a polynomial assumes an infinite zero (designated by $a/0$ where $a \neq 0$) whenever its leading coefficient becomes zero. This convention is consistent with the result obtained using limits (Section 3–11) since at least one zero of $p(x)$ increases without bound as the leading coefficient of $p(x)$ approaches zero. Under the above convention for infinite zeros, we have

THEOREM 4–7. *Given a polynomial $p(x) = a_0x^m + a_1x^{m-1} + \cdots + a_m$, we may immediately write down a polynomial $h(u) = a_0 + a_1u + a_2u^2 + \cdots + a_mu^m$ with zeros that are the reciprocals of those of $p(x)$.*

The question of infinite zeros mentioned above arises, for example, when we take $p(x) = x^2 - x$ with zeros 1 and 0 and apply the above theorem to obtain $h(u) = 0 \cdot u^2 - u + 1$ with zeros 1 and $\frac{1}{0}$. We shall leave the proof of Theorem 4–7 as an exercise (Exercise 6).

We may now perform three transformations upon the zeros of any given polynomial $p(x)$ with zeros $r_1, r_2, \ldots, r_m$. We may find $f(v)$ with zeros $r_j - b$ (Section 4–3), $q(y)$ with zeros ar_j $(a \neq 0)$ (Theorem 4–6), and $h(u)$ with zeros $1/r_j$ (Theorem 4–7). These three transformations are sufficient to prove the following theorem:

THEOREM 4–8. *Given a polynomial $p(x)$ with complex coefficients and zeros r_j $(j = 1, 2, \ldots, m)$, we may obtain a polynomial $q(y)$ with zeros $s_j = (ar_j + b)/(cr_j + d)$ $(j = 1, 2, \ldots, m)$, where a, b, c, d are any complex numbers such that $ad - bc \neq 0$, by using only the coefficients of $p(x)$ and the four fundamental operations.*

The condition $ad - bc \neq 0$ makes it possible (Exercise 7, Section 3–10) to express r_j rationally in terms of s_j, $r_j = (a's_j + b')/(c's_j + d')$, that is, to perform any linear birational transformation upon the zeros of a given polynomial $p(x)$.

We divide the proof of Theorem 4–8 into two cases. When $c = 0$, we use the two successive transformations $t_1 = ax/d$ and $y = t_1 + b/d$. When $c \neq 0$, we use the five transformations $t_1 = cx$,

$$t_2 = t_1 + d = cx + d, \quad t_3 = 1/t_2 = 1/(cx + d),$$
$$t_4 = (b - ad/c)t_3 = (bc - ad)/[c(cx + d)],$$

and $y = t_4 + a/c = (ax + b)/(cx + d)$. Since each of these transformations is of one of the three types that we can perform, our proof of Theorem 4–8 is complete.

One of the most useful applications of the elementary symmetric polynomials is expressed in the following theorem, which we shall prove using transformations of the roots of a polynomial equation.

THEOREM 4–9. *Every rational root of a polynomial equation $a_0x^m + a_1x^{m-1} + a_2x^{m-2} + \cdots + a_m = 0$ with integral coefficients may be expressed in the form b_m/b_0, where $(b_m, b_0) = 1$, b_m divides a_m and b_0 divides a_0.*

This theorem gives a basis for finding all rational roots of any given polynomial equation with integral coefficients in a finite number of steps. It enables us to use at most the quadratic formula to solve completely any polynomial equation with integral coefficients having at most two nonrational roots. For example, the only possible rational roots of $x^3 - x^2 + x - 1 = 0$ are 1 and -1. These may be checked by substitution or by synthetic division. The given polynomial equation may then be expressed as $(x - 1)(x^2 + 1) = 0$, whence its roots are $1, i, -i$.

Theorem 4–9 may, as mentioned above, be proved using transformations of the roots. Suppose that $p(x) = 0$ has a rational root b_m/b_0. We may assume that $(b_m, b_0) = 1$. If we apply Theorem 4–6 and multiply the roots of $p(x) = 0$ by b_0, then S_m and $b_0^m a_m/a_0$ have the same numerical value, the root b_m must divide S_m and therefore b_m must divide $b_0^m a_m$. Since $(b_m, b_0) = 1$, b_m divides a_m, by Theorem 2–9. Similarly, b_0 divides a_0, using Theorem 4–7 (Exercise 10). Since any polynomial with rational coefficients has an associated polynomial with integral coefficients, Theorem 4–9 may also be used to find all rational roots of any polynomial equation with rational coefficients.

The elementary symmetric polynomials have considerable theoretical importance aside from the above-mentioned applications to the transformations of the zeros of polynomials and the solution of polynomial equations with rational coefficients. The elementary symmetric polynomials in the roots of a polynomial equation can always be expressed rationally in terms of the coefficients of the original polynomial.

A polynomial $p(r_1, r_2, \ldots, r_m)$ is said to be symmetric if it is unchanged by every possible interchange of r_j and r_i. For example, $x + y$, xy, $x^2 + y^2$, $xy - x - y$, and $x^3 - xy + y^3 - 2$ are symmetric polynomials in x and y. It can be shown [49; 264] that any symmetric polynomial in the zeros of a polynomial $p(x)$ is a polynomial in the elementary symmetric polynomials and can therefore be expressed rationally in terms of the coefficients of $p(x)$. Exercises 14 and 15 are examples of this property.

The subject of symmetric polynomials or symmetric functions is extensively treated in many texts on the theory of equations. We conclude our brief discussion of this topic with the following exercises. In the next two optional sections we shall return to the problem of solving equations and in particular to the solution of cubic and quartic equations. After that we shall consider methods for determining the number of real roots of a polynomial equation.

EXERCISES

1. Given $x^3 - 2x^2 - 5x + 6 = 0$ with roots $1, -2, 3$, find a polynomial equation with roots $-1, 2, -3$.

2. Generalize the method used in Exercise 1 and give a procedure for finding a polynomial equation $q(y) = 0$ with roots $-r_j$ corresponding to any given $p(x) = 0$ with roots r_j.

3. Given a polynomial $p(x)$, prove Theorem 4–6 by solving for x in the relation $y = kx$ and substituting in $p(x)$.

4. Find a polynomial $q(y)$ with zeros 3 times those of

$$p(x) = x^3 - 3x^2 + 2x - 1.$$

First write down the answer according to Theorem 4–6 and then check this answer by the method given in Exercise 3.

5. Find a polynomial $q(y)$ with zeros -2 times those of

$$p(x) = x^4 + 2x^3 - x^2 + x + 1.$$

Use the same procedure as in Exercise 4.

6. Prove Theorem 4–7, using symmetric polynomials.

7. Rephrase and work Exercise 3 for Theorem 4–7.

8. Write down a polynomial with zeros that are the reciprocals of those of the polynomial $p(x)$ given in Exercise 5. Check this answer by the method given in Exercise 7.

9. Rephrase and work Exercise 3 for Theorem 4–8.

10. Prove that b_0 divides a_0 in Theorem 4–9.

11. Find the rational roots and then solve completely:
   (a) $2y^3 - y^2 - 4y + 2 = 0$,
   (b) $2x^4 - 12x^3 + 19x^2 - 6x + 9 = 0$.

12. Find all the rational roots of
   (a) $3y^4 - 40y^3 + 130y^2 - 120y + 27 = 0$,
   (b) $3y^3 - 2y^2 + 9y - 6 = 0$,
   (c) $108y^3 - 270y^2 - 42y + 1 = 0$,
   (d) $24y^3 - 2y^2 - 5y + 1 = 0$.

13. Find the integral roots of the equations
   (a) $x^4 + 6x^3 + x^2 - 24x - 20 = 0$,
   (b) $x^4 + 11x^3 + 41x^2 + 61x + 30 = 0$.

14. Given a cubic equation $x^3 + ax^2 + bx + c = 0$ with roots r, s, t, find a formula for the symmetric function $r^2 + s^2 + t^2$ in terms of the coefficients of the given equation. Rewrite this formula in terms of the elementary symmetric polynomials S_1, S_2, and S_3.

15. Given a quadratic equation $x^2 + px + q = 0$ with roots r and s, express the following symmetric polynomials as polynomials in the elementary symmetric polynomials

   (a) $r^2 + s^2$, (c) $r^3 - rs + s^3 - 2$,
   (b) $r - rs + s$, (d) $r(r^2 + s - r) + s(s^2 + r - s)$.

***4–9 Cubic equations.** We have stated (Section 4–5) that the general polynomial equations of degrees 1, 2, 3, and 4 may be solved using the four fundamental operations (addition, subtraction, multiplication, and division) and radicals. The linear and quadratic equations were treated in Section 4–5; cubic equations will be treated in this section; quartic equations in the next section. As observed in Section 4–5, there do not exist any similar methods for solving general equations of degree greater than four. We may solve any given polynomial equation with integral coefficients having at most four nonrational roots no matter how large the degree of the equation may be (Section 4–8). We may also solve certain equations of degree greater than four (Section 4–13). However, we still cannot solve in terms of a finite number of rational operations and radicals a general equation $a_0x^m + a_1x^{m-1} + \cdots + a_m = 0$ when m is greater than four.

We shall base our method of solving cubic and quartic equations upon transformations of the roots (Section 4–8). First let us try this method upon the general quadratic equation. The general quadratic equation $ax^2 + bx + c = 0$, $a \neq 0$ with roots r_1, r_2, could have been solved by making the following two transformations upon the roots r_1, r_2. The equation $g(y) = y^2 + 2by + 4ac = 0$ has roots $2ar_j$, by Theorem 4–6. This is similar to the previous derivation of the roots in that the initial is now 1, and we are prepared to divide the coefficient of the linear term by 2. Since the sum of the roots of $g(y)$ is $-2b$, we diminish the roots by $-b$ to obtain a new equation with no linear term, which therefore may be solved by factoring as the difference of two squares. The new equation may be determined by synthetic division (Section 4–3),

$$\begin{array}{llll} 1 & 2b & 4ac & \lfloor -b \\ 0 & -b & -b^2 & \\ \hline 1 & b & 4ac - b^2 & \\ 0 & -b & & \\ \hline 1 & 0 & & \\ 0 & & & \\ \hline 1 & & & \end{array}$$

to be $z^2 + (4ac - b^2) = 0$ with roots s_1, s_2, where $s_j = 2ar_j + b$. Then $s_j = \pm\sqrt{b^2 - 4ac}$ and the relation $r_j = (s_j - b)/2a$ gives the roots r_j in the usual form. A similar procedure may be used to facilitate the solution of cubic and quartic equations.

Let the general cubic equation

$$p(x) = ax^3 + bx^2 + cx + d = 0,\ a \neq 0$$

have roots r_1, r_2, r_3. In order to have the initial 1 and the sum of the roots easily divisible by 3, we use Theorem 4–6 to obtain, after division by a,

$$g(y) = y^3 + 3by^2 + 9acy + 27a^2d = 0,$$

with roots $3ar_j$. As before, we next reduce the roots by $-b$, using synthetic division,

				$\lfloor -b$
1	$3b$	$9ac$	$27a^2d$	
0	$-b$	$-2b^2$	$2b^3 - 9abc$	
1	$2b$	$9ac - 2b^2$	$2b^3 - 9abc + 27a^2d$	
0	$-b$	$-b^2$		
1	b	$9ac - 3b^2$		
0	$-b$			
1	0			
0				
1				

to obtain

$$h(z) = z^3 + (9ac - 3b^2)z + (2b^3 - 9abc + 27a^2d) = 0.$$

This equation is called the *reduced cubic.* We rewrite it in the form

$$z^3 + pz + q = 0,$$

where $p = 9ac - 3b^2$ and $q = 2b^3 - 9abc + 27a^2d$.

There are two well-known procedures for continuing to solve the cubic. We can reduce the problem to simpler form by making a further transformation $z = t - p/3t$ [47; 105], or we can simplify it in a different way by means of the substitution $z = u + v$ [49; 85]. Let us consider the second approach. We replace the single variable z by two variables u, v. These two variables may be required to satisfy another condition which we shall select shortly. Under the substitution $z = u + v$, the above reduced cubic has the form

$$u^3 + v^3 + (3uv + p)(u + v) + q = 0.$$

We now take $3uv + p = 0$ as the new condition on the variables u, v so that the above equation will have the form

$$u^3 + v^3 + q = 0.$$

Then the solution of the reduced cubic is equivalent to the simultaneous solution of $u^3 + v^3 = -q$ and $uv = -p/3$. The cube of the last equation is $u^3v^3 = -p^3/27$. Thus u^3 and v^3 are two variables whose sum is $-q$ and whose product is $-p^3/27$, that is, they are the two roots of

$$t^2 + qt - p^3/27 = 0. \tag{4–6}$$

Therefore we may choose

$$u^3 = -\frac{q}{2} + \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} = A,$$

$$v^3 = -\frac{q}{2} - \sqrt{\frac{q^2}{4} + \frac{p^3}{27}} = B.$$

The possible values of u are $\sqrt[3]{A}, \omega\sqrt[3]{A}, \omega^2\sqrt[3]{A}$, and those of v are $\sqrt[3]{B}, \omega\sqrt[3]{B}, \omega^2\sqrt[3]{B}$, where ω is a primitive cube root of unity (Section 1–17). Since by hypothesis $uv = -p/3$, these values must be associated in pairs as follows:

$$u_1 = \sqrt[3]{A}, \qquad v_1 = \sqrt[3]{B},$$
$$u_2 = \omega\sqrt[3]{A}, \qquad v_2 = \omega^2\sqrt[3]{B},$$
$$u_3 = \omega^2\sqrt[3]{A}, \qquad v_3 = \omega\sqrt[3]{B}.$$

Then, retracing our steps, we see that the roots r_j of the original equation must be given by $r_j = (u_j + v_j - b)/3a$ $(j = 1, 2, 3)$. These formulas for the roots of a cubic equation are known as *Cardan's formulas.*

Thus we have used complex numbers and radicals to express the roots of a general cubic equation in terms of its coefficients. The cubic equation has one real and two conjugate imaginary roots when (4–6) has real distinct roots; three real roots of which at least two are equal when (4–6) has equal roots; three distinct real roots when (4–6) has imaginary roots. These facts and special methods for treating various types of cubic equations are discussed in most texts in the theory of equations. In particular, when the cubic equation has real coefficients and three distinct real roots, it is often convenient to express the roots as real functions of the coefficients using trigonometric functions. This method is especially useful when the cubic equation has rational coefficients and three irrational roots, since in this case [49; 91–92] it is not possible to express any root in terms of real radicals and it is not possible to obtain the roots from Cardan's formulas by rational operations.

Let us now apply the above theory to a simple cubic equation, say $x^3 - x^2 + x - 1 = 0$, that we can also solve by the methods of Section 4–8. The above equation has roots 1, i, and $-i$. Thus the use of Cardan's formulas here becomes merely an illustration of the method for this particular cubic equation. We first multiply the roots by $3a = 3$ and obtain $g(y) = y^3 - 3y^2 + 9y - 27 = 0$. Next we reduce the roots by $-b = 1$, using synthetic division, and obtain

$h(z) = z^3 + 6z - 20$. Now we take $z = u + v$, where $uv = -\frac{6}{3} = -2$. Then $u^3v^3 = -8$ and, by substitution in $h(z)$, $u^3 + v^3 = 20$. These elementary symmetric polynomials in u^3 and v^3 may be used to form a quadratic equation $t^2 - 20t - 8 = 0$ having u^3 and v^3 as its roots. Since the roots are $10 \pm 6\sqrt{3}$, we may take $u^3 = 10 + 6\sqrt{3}$ and $v^3 = 10 - 6\sqrt{3}$. We now know from this method that the given equation has one real and two conjugate imaginary roots. Then, using $\omega = (-1 + i\sqrt{3})/2$, we have

$$u_1 = \sqrt[3]{10 + 6\sqrt{3}}, \qquad v_1 = \sqrt[3]{10 - 6\sqrt{3}},$$
$$u_2 = [(-1 + i\sqrt{3})\sqrt[3]{10 + 6\sqrt{3}}]/2, \qquad v_2 = [(-1 - i\sqrt{3})\sqrt[3]{10 - 6\sqrt{3}}]/2,$$
$$u_3 = [(-1 - i\sqrt{3})\sqrt[3]{10 + 6\sqrt{3}}]/2, \qquad v_3 = [(-1 + i\sqrt{3})\sqrt[3]{10 - 6\sqrt{3}}]/2.$$

Finally, using $r_j = (u_j + v_j + 1)/3$ and a considerable amount of arithmetical simplification, we have $r_1 = 1$, $r_2 = i$ and $r_3 = -i$. The above cubic equation can also be solved by considering rational roots. Cardan's formulas give the same results as the other methods but require more work. We shall resort to the formulas only when other methods fail. Their importance lies in the fact that they give a sure, even though tedious, method for expressing the roots of any cubic equation with complex coefficients in terms of radicals.

EXERCISES

Solve the following equations:

1. $x^3 - 7x^2 + 15x - 9 = 0$
2. $x^3 + 2x + 20 = 0$
3. $x^3 - 3x^2 - 2x + 5 = 0$
4. $24y^3 - 2y^2 - 5y + 1 = 0$

***4–10 Quartic equations.** We now consider the solution of polynomial equations of the fourth degree. As in the case of cubic equations, our method will be based upon transformations of the roots. Also as in the case of the cubic equations, we shall use this method only when other methods, such as finding rational roots (Section 4–8) and multiple roots (Section 4–13), fail.

Let the general quartic equation

$$f(x) = ax^4 + bx^3 + cx^2 + dx + e = 0, \quad a \neq 0$$

have roots r_j $(j = 1, 2, 3, 4)$. The first general method of solving quartic equations was discovered by Ferrari, a pupil of Cardan. Uspensky [49; 94–97] gives a thorough and readable discussion of Ferrari's method. We shali use the method of Descartes [47; 114–117]. We

first multiply the roots by $4a$ and divide the new equation by a to obtain

$$g(y) = y^4 + 4by^3 + 16acy^2 + 64a^2dy + 256a^3e = 0.$$

Then the roots are reduced by $-b$ and an equation of the form

$$z^4 + pz^2 + qz + r = 0 \tag{4–7}$$

is obtained, where p, q, r are polynomials in the coefficients of $f(x)$.

Since in the set of polynomials with real coefficients every irreducible polynomial is linear or quadratic (Section 4–6), the equation (4–7) can be expressed as the product of two quadratic polynomials. If (4–7) has real coefficients, the new polynomials will have real coefficients. Also, since the sum of the roots of (4–7) is zero, the sum of all four roots of the quadratic factors must be zero and the sums for the individual quadratics must be the same except for sign. Thus equation (4–7) may be expressed in the form

$$(z^2 - kz + n)(z^2 + kz + m) = 0. \tag{4–8}$$

Since the left members of (4–7) and (4–8) are identically equal, we have $n + m - k^2 = p$, $k(n - m) = q$, and $nm = r$.

We next eliminate n and m from these equations. We have:

$$k^2(n + m)^2 = k^2(k^2 + p)^2,$$
$$k^2(n - m)^2 = q^2,$$

and by subtraction obtain $4k^2nm = k^2(k^4 + 2k^2p + p^2) - q^2$ or

$$k^6 + 2pk^4 + (p^2 - 4r)k^2 - q^2 = 0,$$

the *resolvent cubic* in k^2. The resolvent cubic can be solved and its roots designated by the complex numbers $4A^2$, $4B^2$, $4C^2$. Then the product of the roots is $q^2 = 64A^2B^2C^2$, and we may choose the A, B, C such that $q = -8ABC$. Similarly, the sum of the roots is $-2p = 4(A^2 + B^2 + C^2)$. If any root of the resolvent cubic is different from zero, suppose $2A \neq 0$. In any case, let $k = 2A$. Then from $n + m - k^2 = p$, $p = -2(A^2 + B^2 + C^2)$, and $k^2 = 4A^2$, we have

$$n + m = k^2 + p = 2(A^2 - B^2 - C^2).$$

Similarly, from $k = 2A$, $q = -8ABC$, and $k(n - m) = q$, we have

$$n - m = -4BC.$$

From these two relations in n and m, we have

$$n = (A + B + C)(A - B - C),$$
$$m = (-A + B - C)(-A - B + C).$$

Since the sums of the factors of n and m are respectively k and $-k$, these four factors are the roots of (4–8) and therefore of (4–7). The roots of the given quartic equation are obtained from those of (4–7) by adding $-b$ and dividing by $4a$. Thus the roots of a general quartic equation can be expressed in terms of the coefficients using radicals. As mentioned in Section 4–5, this process cannot be continued any further, since a general equation of degree greater than four cannot be solved using the four fundamental operations and radicals.

EXERCISES

Solve the following equations:

1. $x^4 - 2x^3 - 3x^2 + 4x + 4 = 0$
2. $3x^4 - 3x^2 - 2x + 5 = 0$
3. $x^4 - x^2 + 10x - 4 = 0$
4. $x^4 - 6x^2 - 8x - 3 = 0$

4–11 Descartes' Rule of Signs. We now return to the task of counting various types of roots of polynomial equations. Any polynomial of degree m with complex coefficients has m complex roots (Section 4–4). The rational roots of any polynomial equation with integral (or rational) coefficients may be found in a finite number of steps (Section 4–8). In this section we shall consider a method for estimating the number of real roots of any given polynomial equation with real coefficients. To be exact, we shall estimate the number of positive roots, count the number of zero roots, and estimate the number of negative roots. In the next section we shall consider a method for determining exactly the number of real roots of a polynomial equation with real coefficients on any interval (Section 3–10) of the form $a < x \leqq b$. The importance of the method in the present section lies in its simplicity.

When all the roots of a polynomial equation $f(x) = 0$ are real and positive, all the elementary symmetric polynomials are positive, and the coefficients of $f(x)$ alternate signs, since

$$\begin{aligned} f(x) &= a_0x^n + a_1x^{n-1} + a_2x^{n-2} + a_3x^{n-3} + \cdots + a_n \\ &= a_0[x^n - S_1x^{n-1} + S_2x^{n-2} - S_3x^{n-3} + \cdots + (-1)^nS_n]. \end{aligned}$$

On the other hand, all the coefficients of $f(x)$ have the same sign when all the roots are negative. We now seek more exact relations between the coefficients of the equation $f(x) = 0$ and the location of its roots.

Two consecutive terms of a real polynomial in which the terms with zero coefficients have been omitted are said to present a *variation*, or a *permanence*, of sign according as their coefficients have unlike or like signs. For example, $x^2 - 1$ has one variation and $x^2 + 1$ has one permanence.

Consider a polynomial $f(x)$ and suppose, for example, that it has no zero coefficients and that the signs of its coefficients are given as in (4–9) below. We next compute the signs of $g(x) = (x - r)f(x)$, where r is any positive number, and indicate the signs of the coefficients of $g(x)$ by $\pm$ whenever the signs depend upon the values of r and the coefficients of $f(x)$.

(4–9)

$f(x)$:	+	+	−	−	−	+	−	−	+	+	−	
$(x - r)$:	+	−										
	+	+	−	−	−	+	−	−	+	+	−	
		−	−	+	+	+	−	+	+	−	−	+
$g(x)$:	+	±	−	±	±	+	−	±	+	±	−	+

The sequence of coefficients of the polynomial $f(x)$ above has five variations in sign; that of $g(x)$ has at least six variations and may have eight but never seven.

In general, for any given polynomial $f(x)$ with real coefficients and any given positive number r, it can be shown [19; 446–447] that the sequence of coefficients of the polynomial $g(x) = (x - r)f(x)$ has at least one more variation in sign than that of $f(x)$. This statement may be proved using the relations $c_0 = b_0$ and $c_j = b_j + c_{j-1}r$, where the c_j are the coefficients of $f(x)$ and the b_j are the coefficients of $g(x)$ (Section 4–2). For example, if the sequence $b_0, b_1, \ldots, b_s$ does not contain any variations, then the sequence $c_0, c_1, \ldots, c_s$ does not contain any variations. If, in addition, the sequence $b_s, \ldots, b_{s+k}$ contains one variation, then the sequence $c_s, \ldots, c_{s+k}$ contains at most one variation. Furthermore, if b_v, b_w is the jth variation in the sequence of coefficients of $g(x)$ and c_t, c_u is the jth variation in the sequence of coefficients of $f(x)$, then $w \leqq u$ (Exercise 6). Thus the sequence of b_j has at least as many variations as the sequence of c_j. Finally, since $b_0 = c_0$ and $g(0) = -rf(0)$, the sequence of b_j [the coefficients of $g(x)$] has more variations than the sequence of c_j [the coefficients of $f(x)$].

If $f(x) = 0$ has $r_1, r_2, \ldots, r_k$ as its positive roots, then

$$f(x) = (x - r_1)(x - r_2) \cdots (x - r_k)Q(x).$$

By repeated applications of the above statement, the sequence of coefficients of $f(x)$ has at least k more variations than that of $Q(x)$, i.e., the number of variations of sign in $f(x)$ is $\geqq k$. Thus a real polynomial equation with V variations in the signs of the coefficients has at most V positive roots.

Now if we write $f(x)$ in the form

$$f(x) = a_0x^n + a_1x^{n-1} + \cdots + a_sx^{n-s},$$

where $0 \leqq s \leqq n$, we may assume $a_0a_s \neq 0$. Then

$$f(x) = x^{n-s}(a_0x^s + a_1x^{s-1} + \cdots + a_s) = x^{n-s}g(x),$$

and the positive roots of $g(x) = 0$ are precisely those of $f(x) = 0$. Since $g(x)$ is continuous for all x, the equation $g(x) = 0$ has an even or odd number, N, of positive roots (not necessarily distinct) according as $0 < a_0a_s$ or $a_0a_s < 0$ (Section 3–13, Exercise 5). Also, V is even if $0 < a_0a_s$, odd if $a_0a_s < 0$. Thus N and V are either both even or both odd, i.e.,

THEOREM 4–10. DESCARTES' RULE OF SIGNS. *A real polynomial with V variations in the signs of its coefficients has $V - 2k$ positive (real) roots where k is a nonnegative integer.*

Since the negative roots of $f(x) = 0$ are equal to the positive roots of $f(-x) = 0$ except for sign (Section 4–8), we also have: *A real polynomial $f(x)$ has $W - 2k$ negative roots where k is a nonnegative integer and W is the number of variations in the signs of the coefficients of $f(-x)$.*

Let us consider a few examples of these two aspects of Descartes' Rule of Signs. The polynomial equation $x^3 - x^2 + 1 = 0$ has either two positive and one negative root or no positive, one negative, and two complex roots. The polynomial equation $x^3 - x^2 + x - 1 = 0$ has either three positive or one positive and two complex roots. The roots of the polynomial equation $x^5 + x^4 - 7x^3 + 5x^2 - x\sqrt{2} + 11 = 0$ fall in one of the following three cases: four positive, one negative, and no complex roots; two positive, one negative, and two complex roots; no positive, one negative, and four complex roots. In the case of the polynomial $x^3 - x^2 + x - 1 = (x - 1)(x^2 + 1)$, we may find the zeros $1, i, -i$ and thereby determine which of the stated cases holds. For any polynomial equation with real coefficients, an exact determination of the proper case may be made using Sturm's Theorem (Section 4–12) without finding the roots.

Descartes' Rule of Signs may also be used to give bounds for the real roots of any polynomial equation with real coefficients, i.e., a *real polynomial equation.* Suppose $f(x) = (x - p)Q(x) + R, 0 < p$ and $Q(x)$ has no variations in the signs of its coefficients. Then the equation $Q(x) = 0$ has no positive root. If also R has the same sign as the coefficients of Q, then at $x = p, f(x) = R$, and for $x > p$, $Q(x)$ and $f(x)$ have the same sign as R, that is, $f(x)$ has no positive zeros greater than p. This test for an upper bound p is most easily applied using synthetic division, where the coefficients of $Q(x)$ and R constitute the third row. The test then becomes: *if p is a positive number such that in the synthetic division of $f(x)$ by $x - p$ all numbers on the third row have the same sign, or are zero, then p is an upper bound for the real zeros of $f(x)$.* A lower bound for the real zeros may be similarly determined as $-q$, where q is an upper bound for the zeros of $f(-x)$.

For example, if $f(x) = 2x^4 - 3x^3 - x^2 - 25x + 30$ and $p = 4$, we have

2	−3	−1	−25	30	| 4
0	8	20	76	204	
2	5	19	51	234	

whence $f(x)$ has no positive zero greater than 4. Similarly, $f(x)$ has no negative zero less than -1, since $f(-x) = 2x^4 + 3x^3 - x^2 + 25x + 30$ has no positive zero greater than 1 from

2	3	−1	25	30	| 1
0	2	5	4	29	
2	5	4	29	59	

There are two other common methods [1; 162–166] of determining upper bounds for the real roots of a real polynomial equation

$$p(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n = 0, \quad a_0 > 0.$$

If $a_j \geqq 0$ $(j = 1, 2, \ldots, k - 1)$, $a_k < 0$, and all negative coefficients are less than or equal to A in absolute value, then $1 + \sqrt[k]{A/a_0}$ is an upper bound for the real roots of $p(x)$. If the absolute value of each negative a_j is divided by the sum of all positive a_i $(i < j)$ and B is the greatest quotient so obtained, then $1 + B$ is an upper bound for the real zeros of $p(x)$.

More exact general information regarding the location of the zeros of any real polynomial can be found from Sturm's Theorem (Section 4–12), which gives the exact number of distinct roots on any interval $a < x \leqq b$.

EXERCISES

1. Discuss the nature of the roots of the following equations:
   (a) $x^6 + 3x^4 + 4x^2 + 2x - 6 = 0$,
   (b) $x^4 - 15x^2 + 7x - 11 = 0$,
   (c) $x^n - 1 = 0$ when n is odd, even,
   (d) $x^n + 1 = 0$ when n is odd, even.
2. Give upper and lower bounds for the real roots of the following equations:
   (a) $x^3 + 2x + 20 = 0$,
   (b) $x^3 - 3x^2 - 2x + 5 = 0$,
   (c) $3x^4 - 6x^2 + 8x - 3 = 0$.
3. Are the bounds given in your answer for Exercise 2 the best possible real bounds, the best possible integral bounds?
4. Prove that the number of negative roots of $f(x) = 0$ is of the form $P - 2k$, where P is the number of permanences of sign in $f(x)$, k is a non-negative integer, and $f(x)$ has no zero coefficients.
5. Prove that if in the synthetic division of $f(x)$ by $x - b$, $b < 0$, the signs of the terms in the third row, associating a suitable sign with any zero terms, can be made to alternate by a suitable choice of b, and if $f(x)$ is of even degree, then the equation $f(x) = 0$ has no negative root less than b.
6. Write out a complete proof of Theorem 4–10.

4–12 Sturm's Theorem. We now consider a method for determining exactly the number of distinct real zeros of any given real polynomial on any interval $a < x \leqq b$. In the next section this method will be extended to determine the number of zeros of any given multiplicity k. Then we shall be able to determine the exact number (including multiplicities) of real roots of any given real polynomial equation $p(x) = 0$.

Given any real polynomial $f(x)$, we may take $f_0 = f(x)$ and $f_1 = cf'(x)$, where $f'(x)$ is the derivative of $f(x)$ (Section 3–14) and c is any positive constant, usually the reciprocal of the greatest common divisor of the coefficients of $f'(x)$. We then apply the Euclidean Algorithm (Section 3–7) to f_0 and f_1 with the modification that the sign of each remainder is changed, i.e., we take

$$\begin{aligned} f_0 &= q_1 f_1 - c_2 f_2, \\ f_1 &= q_2 f_2 - c_3 f_3, \\ &\cdot \\ &\cdot \\ &\cdot \\ f_{k-2} &= q_{k-1} f_{k-1} - c_k f_k, \\ f_{k-1} &= q_k f_k. \end{aligned}$$

Since the signs of the functions will have meaning, only positive factors can be arbitrarily inserted or removed. The array in Section 3–7 can be modified to give

$$\begin{array}{cccccccc} & q_1 & q_2 & q_3 & \cdots & q_k & & \\ f_0 & f_1 & f_2 & f_3 & \cdots & f_k & 0 & \\ q_1 f_1 & q_2 f_2 & q_3 f_3 & q_4 f_4 & \cdots & & & \\ -c_2 f_2 & -c_3 f_3 & -c_4 f_4 & -c_5 f_5 & \cdots & & & \end{array}$$

where the c's are arbitrary positive constants. Then f_k is a greatest common divisor of f_0 and f_1 and a common divisor of the f_j $(0 \leqq j \leqq k)$. The polynomial f_k is not necessarily *the* greatest common divisor since its initial may not be $+1$.

If $f_0 = x^4 - 24x^2 + 16x + 12$, we may take $f_1 = x^3 - 12x + 4$, $f_2 = x^2 - x - 1$, $f_3 = 2x - 1$, and $f_4 = 1$. In this case the polynomials f_0 and f_1 are relatively prime. If $f_0 = x^6 - 3x^2 - 2$, we may take $f_1 = x^5 - x$ and $f_2 = x^2 + 1$. In this case f_3 is identically zero and f_2 is the greatest common divisor of f_0 and f_1. In general, the sequence

$$f_0, f_1, f_2, \ldots, f_k,$$

where f_k is not identically zero, is called the *division sequence* of f_0 and f_1. We also form a sequence of polynomials g_j, letting $g_j = f_j$ if f_k is constant, $g_j f_k = f_j$ if f_k is of positive degree. The sequence

$$g_0, g_1, g_2, \ldots, g_k$$

is called the *Sturm sequence* of f_0. The g's are called *Sturm functions* or *Sturm polynomials* of f_0. For example, the Sturm sequence of the above polynomial $f_0 = x^4 - 24x^2 + 16x + 12$ may be taken as

$$\begin{aligned} g_0 &= x^4 - 24x^2 + 16x + 12, \\ g_1 &= x^3 - 12x + 4, \\ g_2 &= x^2 - x - 1, \\ g_3 &= 2x - 1, \\ g_4 &= 1. \end{aligned}$$

The Sturm sequence of the polynomial $f_0 = x^6 - 3x^2 - 2$ considered above may be taken as

$$\begin{aligned} g_0 &= x^4 - x^2 - 2, \\ g_1 &= x^3 - x, \\ g_2 &= 1. \end{aligned}$$

The Sturm sequence of any polynomial $f_0(x)$ with real coefficients has the following properties.

(*i*) If $f_0(x_1) = 0$, then $g_0g_1 < 0$ when $x = x_1^-$ and $g_0g_1 > 0$ when $x = x_1^+$, where x_1^- and x_1^+ indicate respectively a value of x slightly less than x_1 and a value slightly greater than x_1. This property may be easily visualized in terms of the graph of $y = f_0^2$. When f_0 is linear, the equation $y = f_0^2$ has a single double root at $x = x_1$, the graph is a parabola, and $(d/dx)f_0^2 = 2cf_k^2 \cdot g_0g_1$ is negative at x_1^- and positive at x_1^+, where c is a positive constant and f_k is a real polynomial. In general, every zero of f_0^2 has even multiplicity (Section 4–13) and $(d/dx)f_0^2$ has the same properties as in the special case considered above.

(*ii*) If some $g_j(x_2) = 0$ where $j > 0$, then $g_{j-1}g_{j+1} < 0$ at $x = x_2$. This inequality may be obtained from the identity $f_{j-1} = q_jf_j - cf_{j+1}$, since all common zeros of f_{j-1} and f_j or of f_j and f_{j+1} must be zeros of f_k. When expressed in terms of the g's, where $f_j = f_kg_j$, this identity has the form $g_{j-1} = q_jg_j - cg_{j+1}$, where $(g_{j-1}, g_j) = 1 = (g_j, g_{j+1})$. Thus $g_{j-1}g_{j+1} < 0$ whenever $g_j = 0$.

(*iii*) The polynomial g_k is different from zero for all real values of x. This property is an immediate consequence of the definition of g_k as 1 when f_k is not constant and as f_k when f_k is constant.

When expressed in terms of the sequence of signs of the Sturm functions, the first property above indicates that the first two signs of the sequence are either $-+$ at x_1^- and $++$ at x_1^+ or else they are $+-$ at x_1^- and $--$ at x_1^+, where x_1 is any zero of f_0. The reasoning used for the second property indicates that whenever some g_j $(j > 0)$ vanishes, the polynomials g_{j-1} and g_{j+1} have opposite signs, i.e., we have either $-\ 0\ +$ or $+\ 0\ -$. The third property indicates that g_k never changes sign. Intuitively, we may visualize the variations in the signs of the Sturm sequence of $f_0(x)$ as flowing to the left as the real variable x increases. For example, consider the signs of the above Sturm sequence for $f_0 = x^4 - 24x^2 + 16x + 12$ at the indicated values of x as given in the following array:

x	g_0	g_1	g_2	g_3	g_4
-6	$+$	$-$	$+$	$-$	$+$
-4	$-$	$-$	$+$	$-$	$+$
-1	$-$	$+$	$-$	$-$	$+$
0	$+$	$+$	$-$	$-$	$+$
1	$+$	$-$	$-$	$+$	$+$
2	$-$	$-$	$+$	$+$	$+$
4	$-$	$+$	$+$	$+$	$+$
5	$+$	$+$	$+$	$+$	$+$

Formally, the Sturm sequence of $f_0(x)$ will be a sequence of constants for any real value of x, say $x = a$. We shall use the symbol S_a to indicate the number of variations in the sequence when $x = a$. From properties (*ii*) and (*iii*) S_x cannot change as x passes through any zero of g_j $(j > 0)$ which is not a zero of g_0. From (*i*) if $f_0(x_1) = 0$, then (Section 4–13) $g_0(x_1) = 0$, and the sequence has a variation in sign in its first two terms at $x = x_1^-$ but no variation at $x = x_1^+$. Thus S_x changes only at zeros of $f_0(x)$ and decreases by 1 whenever x increases through zeros of $f_0(x)$. Thus we have

THEOREM 4–11. STURM'S THEOREM. *A real polynomial $f(x)$ has exactly $S_a - S_b$ distinct real zeros on the interval $a < x \leqq b$.*

The Sturm sequence of the polynomial $x^4 - 24x^2 + 16x + 12$ is given above. For large negative values of x only the leading terms need to be considered (Section 4–5), and the Sturm functions have signs $+\ -\ +\ -\ +$ with four variations. For large positive values of x, we have $+\ +\ +\ +\ +$ with no variations. Thus the polynomial $x^4 - 24x^2 + 16x + 12$ has four distinct real zeros.

The Sturm sequence g_0, g_1, g_2 of $f_0 = x^6 - 3x^2 - 2$ was obtained above by dividing the polynomials f_0, f_1, f_2 by f_2, since f_2 is not a constant. For large negative values of x their signs are $+\ -\ +$ with two variations. For large positive values of x their signs are $+\ +\ +$ with no variations. Thus the polynomial $x^6 - 3x^2 - 2$ has two distinct real zeros.

Sturm's Theorem may also be used for determining bounds for the roots of polynomial equations and indeed for isolating the roots on arbitrary small intervals. For example, $x^6 - 3x^2 - 2 = 0$ has two distinct real roots as determined above. At $x = 2$ the signs of the corresponding Sturm functions are $+\ +\ +$, the same as for large positive values of x. Thus by Theorem 4–11, the equation $x^6 - 3x^2 - 2 = 0$ has no real roots greater than 2. Similarly, at $x = -2$ the signs are $+\ -\ +$, the same as for large negative values of x, and there are no real roots less than -2. At $x = 1, 0$, and -1 the signs of the Sturm functions that are different from zero, i.e., the signs of the Sturm functions that have signs, are $-\ +$ with one variation. Thus $S_{-2} = 2$, $S_{-1} = 1$, $S_0 = 1$, $S_1 = 1$, and $S_2 = 0$, whence there is one real root satisfying $-2 < x < -1$ and one real root satisfying $1 < x < 2$. The intervals $-2 < x \leqq -1$ and $1 < x \leqq 2$ were replaced by $-2 < x < -1$ and $1 < x < 2$, since $x = -1$ and $x = 2$ are not roots. The process of finding intervals containing the roots could be continued

using $S_{-\frac{1}{2}}, S_{\frac{1}{2}}, \ldots$. In general, we shall say that the real roots of an equation have been *isolated* when an interval $a < x < b$ has been found for each real root such that $b - a \leqq 1$ and the interval contains only one of the distinct real roots.

We may now determine exactly the number of distinct real roots of any given real polynomial equation, i.e., any given polynomial equation $p(x) = 0$ with real coefficients. Also we may isolate each root on an arbitrarily small interval. In the next section we shall define the multiplicity of a root and determine the total number of real roots on any interval, using their multiplicities.

Exercises

1. Test for real roots:
   (a) $x^4 - 6x^3 + 7x^2 + 6x - 2 = 0$ (see Exercise 4, Section 3–7),
   (b) $x^4 - 2x^2 + 12x - 8 = 0$,
   (c) $x^4 - 13x^2 + 4x + 2 = 0$,
   (d) $x^4 - x^2 + 10x - 4 = 0$.
2. Isolate all the real roots by Sturm's Theorem:
   (a) $x^3 + 2x + 20 = 0$,
   (b) $x^3 - 3x^2 - 2x + 5 = 0$,
   (c) $3x^4 - 6x^2 + 8x - 3 = 0$.

4–13 Multiple roots. A number b has been defined (Section 4–1) to be a zero of a polynomial $p(x)$ if and only if $p(b) = 0$. Thus 2 is a zero of $x - 2$, $x^2 - 4x + 4$, $x^2 - 3x + 2$, and $x^3 - 2x^2 + 4x - 8$. This definition and Theorem 4–1 imply that any polynomial $p(x)$ with a zero b may be written in the form $p(x) = (x - b) \cdot q(x)$, where $q(x)$ is a polynomial. In the above examples, $x - 2 = (x - 2) \cdot 1$, $x^2 - 4x + 4 = (x - 2) \cdot (x - 2)$, $x^2 - 3x + 2 = (x - 2) \cdot (x - 1)$, and $x^3 - 2x^2 + 4x - 8 = (x - 2) \cdot (x^2 + 4)$. In other words, b is a root of $p(x) = 0$ if and only if $x - b$ divides $p(x)$. We say that b is a *multiple root* of $p(x) = 0$ if and only if $(x - b)^2$ divides $p(x)$. Thus 0 is a multiple root of $x^4 - 7x^2$.

We also speak of the multiplicity of a root b of $p(x) = 0$ indicating the greatest integral power of $(x - b)$ that divides $p(x)$. Thus, for example, the root 0 has multiplicity 5 in $x^8 - 7x^7 + x^5 = 0$. In general, a root b of $p(x) = 0$ has *multiplicity* k if $(x - b)^k$ divides $p(x)$ and $(x - b)^{k+1}$ does not divide $p(x)$, where k is a positive integer. A root of multiplicity 1 is called a *simple root.*

The calculation of the multiplicity of a root b is often most easily

accomplished using a change of variable to express $p(x)$ in the form $q(x - b)$. This may be done using synthetic division (Section 4–3) or Taylor's formula (Section 3–14). In the notation of Taylor's formula we have the following identity for any polynomial $f(x)$:

$$f(x) = f(a) + f'(a)(x - a) + \cdots + f^{(n)}(a)\frac{(x - a)^n}{n!}.$$

It is now evident that if $f(a) = 0$, then $f(x)$ has $(x - a)$ as a factor; if $f(a) = f'(a) = 0$, then $f(x)$ has $(x - a)^2$ as a factor; and, in general, if $f(a) = f'(a) = \cdots = f^{(k)}(a) = 0$, then $f(x)$ has $(x - a)^{k+1}$ as a factor. Since the coefficients in Taylor's formula are uniquely determined, the converse is also true, i.e., if $f(x)$ has $(x - a)^{k+1}$ as a factor, then $f(a) = f'(a) = \cdots = f^{(k)}(a) = 0$. Furthermore, if $f(x)$ has $(x - a)^{k+1}$ as a factor, then $f'(x)$ has $(x - a)^k$ as a factor, $f''(x)$ has $(x - a)^{k-1}$ as a factor, and so forth. Thus any root of $f(x) = 0$ of multiplicity $m > 1$ is a root of $f'(x) = 0$ of multiplicity $m - 1$. In particular, a simple root of $f(x) = 0$ is not a root of $f'(x) = 0$. If $f(x) = 0$ and $f'(x) = 0$ have a common root r that is a zero of $f'(x)$ of multiplicity $m - 1$, then r is a zero of $f(x)$ of multiplicity m. Finally, if $f_k(x)$ is a greatest common divisor of $f(x)$ and $f'(x)$ and $f = f_k g_k$, then $g_k(x)$ has the same distinct zeros as $f(x)$ but no multiple zeros. Thus in Section 4–12 the zeros of $g_0(x)$ are simple and are precisely the distinct zeros of $f_0(x)$. Also, $f_0(x)$ has multiple zeros if and only if $f_k(x)$ is of positive degree.

Given any interval $a < x \leqq b$ and any real polynomial $f_0(x)$, we may find the division sequence of $f_0(x)$, say, $f_0, f_1, \ldots, f_{k1}$. Since f_{k1} is an associate of the greatest common divisor of f_0 and f_0', f_{k1} is a constant if and only if $f_0(x)$ has only simple zeros. If f_{k1} has positive degree, the Sturm sequence $g_0, g_1, \ldots, g_{k1}$ is called the first Sturm sequence of f_0. In either case, Theorem 4–11 gives the number of distinct real zeros, i.e., the number N_1 of zeros of $f_0(x)$ of multiplicity at least one. We now use the fact that any zero of $f_0(x)$ of multiplicity m is a zero of $f_0'(x)$ and therefore of $f_{k1}(x)$ of multiplicity $m - 1$. Thus if f_{k1} is not a constant, we find its division sequence $f_{k1}, f_{k1}', \ldots f_{k2}$ and its Sturm sequence, the second Sturm sequence of f_0. If $f_{k2} = c(f_{k1}, f_{k1}')$ is constant, $f_0(x)$ has no root of multiplicity greater than two. In any case, Theorem 4–11 for f_{k1} gives the number N_2 of zeros of $f_0(x)$ of multiplicity at least two. If f_{k2} has positive degree, we find its Sturm sequence (the third Sturm sequence of f_0) and the number N_3 of zeros of multiplicity at least three. Since

$f_0(x)$ has finite degree, this process can be repeated only a finite number of times. If N_i is the number of zeros of $f_0(x)$ of multiplicity at least i as obtained from the ith Sturm sequence of f_0, then the number of zeros of multiplicity exactly j is $N_j - N_{j+1}$. In particular, the number of simple zeros is $N_1 - N_2$. This procedure therefore gives *Sturm's Theorem for Multiple Roots* [48] by means of which the number $N_j - N_{j+1}$ of real zeros of $f_0(x)$ on any interval $a < x \leqq b$ and of any multiplicity j may be computed without determining the zeros themselves.

The zeros of $h_1 = f_0/f_{k1}$ are the zeros (real or imaginary) of f_0 of multiplicity at least one; those of $h_2 = f_{k1}/f_{k2}$ the zeros of f_0 of multiplicity at least two; those of $h_j = f_{k(j-1)}/f_{kj}$ the zeros of f_0 of multiplicity at least j. Therefore, the zeros of $s_1 = h_1/h_2$ are precisely the simple zeros of f_0; those of $s_2 = h_2/h_3$ the double zeros; those of $s_j = h_j/h_{j+1}$ the zeros of multiplicity j, where $s_n = h_n$ if f_{kn} is constant. The f_{kj} are polynomials obtained by the greatest common divisor process; the h_j and s_j are polynomials obtained by division. These operations can be carried out whether $f_0(x)$ has real or complex coefficients. Thus for any polynomial equation $f(x) = 0$ and for any positive integer j we may obtain a polynomial equation $s_j = 0$ whose roots are the distinct zeros (real and imaginary) of $f(x)$ of multiplicity j. Then, using these polynomials s_j, we have

THEOREM 4–12. *A polynomial equation $f(x) = 0$ can be solved (i) using the quadratic formula if it has at most two roots of every multiplicity k, (ii) in terms of its coefficients using radicals if it has at most four roots of every multiplicity k.*

Example. The equation

$$f_0(x) = x^{10} - x^8 - 5x^6 + x^4 + 8x^2 + 4 = 0 \tag{4–10}$$

has at most two positive roots and at most two negative roots (Section 4–11). It has no rational roots (Theorem 4–9). Before removing the greatest common divisors to obtain the Sturm sequences, we compute the division sequences:

$$\begin{aligned} f_0 &= x^{10} - x^8 - 5x^6 + x^4 + 8x^2 + 4, \\ f_1 &= 5x^9 - 4x^7 - 15x^5 + 2x^3 + 8x, \\ f_2 &= x^8 + 10x^6 - 3x^4 - 32x^2 - 20, \\ f_3 &= x^7 - 3x^3 - 2x, \\ f_{k1} &= -x^6 + 3x^2 + 2; \end{aligned}$$

$$\begin{aligned} f_{k1} &= -x^6 + 3x^2 + 2, \\ c_1 f'_{k1} &= -x^5 + x, \\ f_{k2} &= -x^2 - 1; \end{aligned}$$

$$\begin{aligned} f_{k2} &= -x^2 - 1, \\ c_2 f'_{k2} &= -x, \\ f_{k3} &= 1; \end{aligned}$$

where f_{k1} is a greatest common divisor (GCD) of f_0 and f'_0, f_{k2} is a GCD of f_{k1} and f'_{k1}, and f_{k3} is a GCD of f_{k2} and f'_{k2} (that is, f_{k2} and f'_{k2} are relatively prime).

The first Sturm sequence of (4–10) is then

$$\begin{aligned} g_{10} &= -x^4 + x^2 + 2 = f_0/f_{k1}, \\ g_{11} &= -5x^3 + 4x = f_1/f_{k1}, \\ g_{12} &= -x^2 - 10 = f_2/f_{k1}, \\ g_{13} &= -x = f_3/f_{k1}, \\ g_{14} &= 1 = f_{k1}/f_{k1}; \end{aligned}$$

the second Sturm sequence is

$$\begin{aligned} g_{20} &= x^4 - x^2 - 2 = f_{k1}/f_{k2}, \\ g_{21} &= x^3 - x = c_1 f'_{k1}/f_{k2}, \\ g_{22} &= 1 = f_{k2}/f_{k2}; \end{aligned}$$

and the third Sturm sequence is

$$\begin{aligned} g_{30} &= -x^2 - 1 = f_{k2}, \\ g_{31} &= -x = c_2 f'_{k2}, \\ g_{32} &= 1 = f_{k3}. \end{aligned}$$

From the first Sturm sequence we see that (4–10) has two distinct real roots, one positive and one negative, since $S_{-\infty} = 3$, $S_0 = 2$, and $S_\infty = 1$. Similarly, from the second Sturm sequence we see that (4–10) has two distinct real roots of multiplicity at least two, one positive and one negative. From the third Sturm sequence we see that there are no real roots of multiplicity greater than two.

In the notation of the paragraph preceding Theorem 4–12, the zeros of

$$h_1 = -x^4 + x^2 + 2 = f_0/f_{k1}$$

are the zeros of f_0 of multiplicity at least one, those of

$$h_2 = x^4 - x^2 - 2 = f_{k1}/f_{k2}$$

are the zeros of f_0 of multiplicity at least two, and those of

$$h_3 = -x^2 - 1 = f_{k2}/f_{k3}$$

are the zeros of f_0 of multiplicity at least three. We note that $h_j = g_{j0}$, since both are defined in the same manner. Continuing in the earlier notation, the zeros of

$$s_1 = -1 = h_1/h_2$$

are the zeros of f_0 of multiplicity exactly one, those of

$$s_2 = -x^2 + 2 = h_2/h_3$$

are the zeros of f_0 of multiplicity exactly two, and those of

$$s_3 = -x^2 - 1 = h_3$$

are the zeros of f_0 of multiplicity exactly three. From s_1 we see that the equation (4–10) has no simple roots, from s_2 it has $\sqrt{2}$ and $-\sqrt{2}$ as double roots, and from s_3 it has i and $-i$ as triple roots. Thus the roots of (4–10) are $\sqrt{2}$, $\sqrt{2}$, $-\sqrt{2}$, $-\sqrt{2}$, i, i, i, $-i$, $-i$, $-i$.

The calculation of the Sturm series of a polynomial is often a long and tedious process. However, the solution of a cubic or quartic equation using the formulas is also a long process (Sections 4–9 and 4–10). Theorem 4–11 and the above methods may be used for any polynomial with complex coefficients regardless of its degree to obtain multiple and also simple roots whenever the s_j may be solved by our previous methods. The s_j have lower degrees than the given polynomial if and only if there are multiple roots. In particular, if all roots are simple roots, then s_1 is an associate of the given polynomial. The methods of this section are important in that they enable us to solve equations that could not be solved by methods previously considered. In the next and final section of this chapter we shall consider methods for approximating the real roots of polynomial equations in one variable with real coefficients.

Exercises

1. Solve:
   (a) $x^4 - 4x^3 - 2x^2 + 12x + 9 = 0$,
   (b) $x^4 - 2x^3 - 3x^2 + 4x + 4 = 0$,
   (c) $x^4 - 9x^3 + 9x^2 + 81x - 162 = 0$,
   (d) $x^4 - 6x^2 - 8x - 3 = 0$,
   (e) $x^3 - 7x^2 + 15x - 9 = 0$.

2. Solve $x^{10} - 5x^8 - 5x^6 + 45x^4 - 108 = 0$.
3. Give a method for solving the equation in Exercise 2 by first finding the rational roots of a related equation.

4–14 Approximate solutions. We conclude our brief study of the theory of equations with a discussion of two methods for approximating the real roots of a polynomial equation $f(x) = 0$ with real coefficients. By Sturm's Theorem or even by trial and error we may determine intervals of the form $n < x \leqq n + 1$ on which the roots occur (Section 4–12). The following procedures represent two methods of taking successive approximations which approach the root as a limit. Since all rational roots may be obtained by Theorem 4–9, approximate methods may be necessary only for irrational roots.

Newton's method. Suppose $f(x) = 0$ has a root $a + h$ where the number a is known, $f'(a) \neq 0$, and h^2 is less than one. Taylor's formula at $x = a + h$ gives

$$f(a + h) = f(a) + f'(a)h + \frac{f''(a)}{1 \cdot 2} h^2 + \cdots = 0,$$

since $a + h$ is a root. For small values of h the terms containing h^2 as a factor are neglected to give $f(a) + f'(a)h = 0$ or $h = -f(a)/f'(a)$ as a first approximation for h, and $a_1 = a - f(a)/f'(a)$ as a first approximation for the root $a + h$. This process is repeated taking $a_2 = a_1 - f(a_1)/f'(a_1)$ and, in general, $a_{j+1} = a_j - f(a_j)/f'(a_j)$. Although a_j must be chosen such that $f'(a_j) \neq 0$, this causes no difficulty, since $f'(x)$ has only a finite number of zeros. Thus if $f'(x) = 0$ for some value of $x = a_j$, we merely replace a_j by a slightly different value at which $f'(x) \neq 0$. The sequence of values $a, a_1, a_2, \ldots, a_j, \ldots$ approaches $a + h$ as a limit that can be approximated to any desired degree of accuracy. For example: $f(x) = x^5 - 3x + 1 = 0$ has a root between zero and one; $f'(x) = 5x^4 - 3$. If $a = 0$, the differences $a_1 - a = \frac{1}{3}$, $a_2 - a_1 = .0014, \ldots$ approach zero very rapidly and the desired root is approximately 0.3346. Newton's method may also be used for any function $f(x)$ which has a first derivative.

The second method of approximation, *Horner's method*, gives the successive digits in the decimal expansion of the root. For example, $x^3 - 12x^2 + 5x - 17 = 0$ can be determined by Sturm's Theorem or by trial and error to have a root between 10 and 20. The first digit of the root may thus be taken as 1. We now use synthetic division, as in Section 4–3, to diminish the roots by 10:

$$
\begin{array}{rrrrl}
1 & -12 & 5 & -17 & \mid 10 \\
0 & 10 & -20 & -150 & \\
\hline
1 & -2 & -15 & -167 & \\
0 & 10 & 80 & & \\
\hline
1 & 8 & 65 & & \\
0 & 10 & & & \\
\hline
1 & 18 & & & \\
0 & & & & \\
\hline
1 & & & &
\end{array}
$$

and seek a root of the new equation $y^3 + 18y^2 + 65y - 167 = 0$ between 0 and 10. This root is found to lie between 1 and 2. Thus the second digit of the root is 1. Accordingly, we diminish the roots by 1 and seek a root of the new equation $z^3 + 21z^2 + 104z - 83 = 0$ between 0 and 1. This root is found to be between .6 and .7, the roots reduced by .6 and a root of the new equation sought between zero and one-tenth. Continuing this process, one may calculate any finite number of decimal places of the desired root, 11.6 . . . . After the first few steps, a useful approximation of the next digit may be obtained by considering only the last two terms of the equation at hand. Negative roots of $f(x) = 0$ may be computed by changing the signs of the positive roots of $f(-x) = 0$.

There are other methods of approximating roots [49; 151–180], as well as rules for reducing the labor in the above procedures. However, enough has been presented to indicate how any real root of a real polynomial equation may be determined to any desired number of decimal places. For further details on this and other topics mentioned in this chapter the reader should consult a textbook in the theory of equations.

In this chapter we have discussed methods for finding the zeros of a polynomial in one variable. We found formulas for determining the roots of polynomial equations of degrees one, two, three, and four (Sections 4–5, 4–9, 4–10) with complex coefficients. We considered polynomial equations of arbitrary degree in terms of their multiple roots and real polynomial equations of arbitrary degree in terms of their rational roots. This study of polynomials and polynomial equations represents a goal toward which we have worked steadily while studying our number system, theory of numbers, and theory of polynomials. However, even though it is a notable step in our study, it is by no means the end of the road. At nearly every step along the way there have been opportunities to specialize and develop interesting concepts. We have considered only a few basic fundamentals relative to the vast field of algebra. The door is now open to a wide variety of topics, of which we shall be able to select only a few.

The remaining three chapters may be read in any order, according to the interest of the reader. Chapter 5 on matrices and determinants along with their applications to linear dependence, solutions of systems of linear equations, and geometric transformations is the most fundamental. Chapter 6 contains an important application of our algebraic theories to classical construction problems in geometry and, in particular, contains a proof of the impossibility of trisecting an arbitrary given angle using only straightedge and compasses. Finally, Chapter 7 is an introduction to the graphical representations of functions corresponding somewhat to our previous development of sets of functions in Chapter 3.

EXERCISES

1. Find the roots of $x^3 - 3x^2 - 2x + 5 = 0$ to four decimal places.
2. Find the real root of $x^3 + 2x + 20 = 0$ to five decimal places.
3. A flag pole stands one hundred feet high and ten feet from a ten foot pole. If the flag pole breaks over (top section does not separate from base section) in such a way that it touches the top of the ten foot pole and just touches the ground, find the height of the break.

CHAPTER 5

DETERMINANTS AND MATRICES

Determinants and matrices have many practical applications in mathematics and other sciences. We shall first take a general view of their historical development (Section 5–1), then define determinants in terms of matrices (Section 5–2) and permutations (Sections 5–3 to 5–6), study a few of their properties (Sections 5–7 to 5–10), and consider several of their applications. In particular, we shall consider the use of determinants and matrices in the solution of systems of linear equations (Sections 5–11 and 5–12), linear dependence (Section 5–13), analytic geometry (Section 5–14), and geometric transformations (Section 5–15).

5–1 Historical development. The first notion of a determinant was probably due to Leibniz in the last part of the seventeenth century. He used symbols similar to our present determinant notation to simplify the expressions arising in the solution of systems of linear equations. For example, consider the following system of linear equations in two variables:

$$a_1x + b_1y = c_1,$$
$$a_2x + b_2y = c_2.$$

If the first equation is multiplied throughout by $+b_2$, the second by $-b_1$, and the two resulting equations added, we obtain $(a_1b_2 - a_2b_1)x = c_1b_2 - c_2b_1$ or, in the notation of determinants,

$$\begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix} x = \begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix},$$

which may be solved for x if $a_1b_2 - a_2b_1 \neq 0$. This procedure was expressed as a formal rule for systems of n linear equations in n variables by Gabriel Cramer in 1750. Accordingly, this method is called Cramer's Rule (Section 5–11). This rule expresses one of the first and basic applications of determinants.

During the two centuries since the formulation of Cramer's Rule, determinants have been used in many ways. A few of these applications will be discussed in the present chapter; many others cannot

be appreciated until the reader has studied the special branch of mathematics in which they are used. Bezout (1799) used determinants in his method of elimination by linear equations. Sylvester (1840) used determinants in his dialytic method of elimination. The work in the theory of determinants of Vandermonde, Jacobi, and others has been recognized by associating their names with special types of determinants. Vandermonde's determinant (Exercise 22, Section 5–9) may be used, for example, in discussing the roots of a cubic equation. Wronskians and jacobians have significance in advanced mathematical theories. Resultants, eliminants, alternants, orthogonal determinants, symmetric determinants, skew-symmetric determinants are a few of the many other types of determinants having special applications in mathematical theories.

Considerable advances in the general theory of determinants were made by Cauchy (1815) and Jacobi (1841). Shortly thereafter the concept of a square array designating a determinant was extended to the quite different concept of a rectangular array called a matrix. The matrix is now the fundamental concept, with numerous theoretical and practical applications. Every square matrix with elements from a ring has an associated determinant. The theory of determinants has become a part of the theory of matrices. We shall restrict our study of matrices to concepts needed in the applications mentioned at the beginning of this chapter. Other applications and further details of the theory may be found in texts such as [9], [16], [39], [44], and [49].

5–2 Matrices. A *matrix* is defined to be a rectangular array such as

$$\begin{bmatrix} 1 & 2 & 3 \\ 5 & 6 & 7 \end{bmatrix},$$

and, in general,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}, \tag{5–1}$$

where the elements a_{ij} may be chosen from any designated set of numbers, polynomials, or elements of a given ring of elements. We shall be primarily concerned with matrices having real numbers or

polynomials as elements. Matrices with elements from an arbitrary ring or field could be considered with very little modification of our discussion.

The matrix (5–1) has m rows and n columns. The notation a_{ij} is chosen so that the first subscript (*row index*) designates the row and the second subscript (*column index*) designates the column in which the element is situated. For example, a_{12} is in the first row and the second column; a_{31} is in the third row and the first column; a_{ij} is in the ith row and the jth column.

When $m = n$, we have a *square matrix* such as

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \qquad \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix},$$

or, in general,

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \dots & a_{3n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{bmatrix}. \tag{5–2}$$

In Section 5–7 we shall use the permutations of the subscripts of the elements and associate with any square matrix a polynomial in the elements of that matrix. This polynomial will be called the *determinant* of the matrix. When the elements of a square matrix are numbers, the determinant of the matrix is also a number.

The determinant of a matrix is defined only for square matrices and may be designated by using straight lines in place of the square brackets used to designate a matrix. For example, the determinant of the matrix

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

may be designated by

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}. \tag{5–3}$$

The matrix (5–1) may also be designated by $[a_{ij}]$, $i = 1, 2, \dots, m$; $j = 1, 2, \dots, n$. The square matrix (5–2) may be designated by $[a_{ij}]$, $i, j = 1, 2, \dots, n$ or simply as $[a_{11}, a_{22}, \dots, a_{nn}]$ in terms of its elements with equal row and column indices, i.e., the elements on its

principal diagonal. The determinant of (5–2) may be designated using straight lines as in (5–3); by $|a_{ij}|$, $i, j = 1, 2, \dots, n$; or by $|a_{11}, a_{22}, \dots, a_{nn}|$. These three notations for matrices and determinants are introduced to facilitate reference to other texts on matrices and determinants. Unfortunately, mathematical notation is not well standardized. However, an understanding of the above should enable the reader to recognize quickly the notation in any other text. For our own purposes we have used straight lines for determinants and square brackets for matrices in all three notations.

We have now defined a matrix and been told that, using permutations, a determinant may be associated with each square matrix. In the next few sections we shall define and discuss some properties of permutations.

Exercises

1. Write down five matrices.
2. Do there exist determinants associated with any of the matrices given in Exercise 1? Indicate these determinants when such exist.
3. Give four square matrices and designate the determinant associated with each.
4. The order of a square matrix (Section 5–7) is equal to the number of elements on its principal diagonal. Indicate the order of each of the matrices given in Exercise 3.
5. Give three square matrices of different orders.
6. Designate the determinant of the matrix $[a_{11}, a_{22}, a_{33}, a_{44}]$ in three ways.
7. Give a matrix of order three having complex numbers as its elements.
8. Give a matrix of order three having polynomials in x of positive degree as its elements.
9. Give matrices having each of the following properties: (a) two rows and three columns, (b) one row and three columns, (c) three rows and one column, (d) one row and one column.

5–3 Permutations. The various orders in which the elements of any given set may be arranged in a row are called the *permutations* of the set. For example, given the integers 1 and 2, we have two permutations 1, 2 and 2, 1; given the integers 1, 2, 3, we have six permutations 123, 132, 213, 231, 321, and 312 where the commas in each permutation have been omitted for convenience. We shall omit such commas whenever it is possible without causing confusion to the reader. However, given two integers such as 11 and 17 we shall not,

of course, be able to omit the commas in writing the two permutations 11, 17 and 17, 11.

Let us consider the number P_n of permutations of a given set of n distinct elements. $P_2 = 2$, since any two elements a and b may be arranged in two orders: ab and ba. A third element c may be introduced into each of these two permutations in three ways: before either element or after both, thus giving $3 \cdot 2 = 6$ permutations of the three elements. We write $P_3 = 3!$. In a similar manner, given any positive integer k, an additional element may be introduced into each of the permutations of k elements in $k + 1$ different ways: before each element and after all of them. Thus if P_k denotes the number of permutations of k elements, then $P_{k+1} = (k + 1)P_k$. Thus $P_4 = 4!$, $P_5 = 5!, \ldots, P_n = n!$ for all positive integral values of n (Exercise 1).

We have now defined a permutation as a linear arrangement such as

1 2 3 4 5 6 7 8 9 10 11 12

as contrasted with a circular (for example as on a clock, Fig. 5–1) or other arrangement of any given set of elements. Next we take one order (permutation) of the given set of elements as their *natural order* and consider all permutations of these elements with reference to their natural order. For example, it is customary to take the natural order of any set of consecutive positive integers as the order used in counting; the natural order of any finite set of real numbers as the order of increasing numerical value; the natural order of any set of letters from an alphabet as their alphabetical order. Thus we shall consider each of the permutations

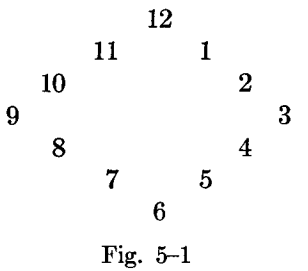


Fig. 5–1

(5–4) 12345, 3567, *acfhm*

to be in their natural order. Similarly, each of the permutations

(5–5) 21345, 3576, *afchm*

is in an order different from its natural order. This difference, or the relation between two permutations such as 12345 and 21345 of a given set of elements, may be expressed in terms of inversions (Section 5–4).

Exercises

1. Give a complete proof by mathematical induction (Section 1–4) that $P_n = n!$ for any positive integer n.
2. List all the permutations of the set of letters *cat*.
3. List the 24 permutations of the set *duck*.

5–4 Inversions. Any two elements, whether adjacent or not, standing in their natural order in a permutation constitute a *permanence;* any two elements standing in an order that is not their natural order constitute an *inversion.* For example, the permutation 1, 2 is called a permanence; 2, 1 an inversion. Given a permutation *daecb* and assuming the alphabetical order to be the natural order, we have permanences *de, ae, ac, ab,* and inversions *da, dc, db, ec, eb, cb.* Given any permutation, we may determine the permanences and inversions as above by considering the first element with each of the other elements, the second element with each of the elements following it, the third element with each of the elements following it, ... In this way we may associate with each permutation a unique nonnegative integer, the number of inversions in the permutation. Thus any given permutation may be classified as *even* or *odd* according as the number of inversions in the permutation is even or odd. Each of the permutations in (5–4) is in its natural order and is an even permutation, since it has no inversions (zero is an even number). Each of the permutations in (5–5) is an odd permutation, since each contains exactly one inversion. Since each permutation of any given set of elements is either even or odd, we shall refer to the *class* of even permutations and the class of odd permutations.

Given the permutation 4132, we may consider the differences $1 - 4$, $3 - 4$, $2 - 4$, $3 - 1$, $2 - 1$, $2 - 3$ and find that four of the differences are negative. The permutation has four inversions and is even. In general, if we associate a positive number with each permanence and a negative number with each inversion, then the product of all such numbers arising from a given permutation is positive if the permutation is even, negative if the permutation is odd. Since an integer k precedes an integer m in their natural order if and only if $m - k$ is positive, a pair of numbers km is a permanence if $m - k$ is positive, an inversion if $m - k$ is negative. Thus given any permutation of numbers, we may consider the sign of the product R of the differences obtained by subtracting each element in the permutation in turn from each of the elements following it. For example, given the permuta-

tion 4132, we take the product of the differences considered above and find that

$$R = (1-4)(3-4)(2-4)(3-1)(2-1)(2-3) \text{ is positive}$$

and, as before, the permutation 4132 is even.

Given the permutation 1432, we find that

$$R = (4-1)(3-1)(2-1)(3-4)(2-4)(2-3) \text{ is negative}$$

and the permutation 1432 is odd. This odd permutation 1432 may be obtained from the even permutation 4132 by interchanging the elements 1 and 4. In general, we shall find (Theorem 5–4) that the interchange of any two elements of a permutation (i.e., a *transposition*) changes the class of the permutation from even to odd or from odd to even. In the next section we shall prove the above result for transpositions of adjacent elements. We shall also prove that any permutation of a given set of elements may be obtained from the set of elements taken in their natural order by a sequence of transpositions of adjacent elements.

Exercises

1. List the inversions in each of the following permutations: 7132, 71452, 635421, 192837465.

2. Classify each of the permutations in Exercise 1 as even or odd (a) in terms of the number of inversions counted, (b) in terms of the sign of the product of the differences R.

3. Indicate the transposition used on each of the three permutations in (5–4) to obtain the corresponding permutations in (5–5).

4. Use the product symbol $\prod$ as in the special case

$$(x_2 - x_1)(x_3 - x_1)(x_3 - x_2) = \prod_{1 \leqq i < j \leqq 3} (x_j - x_i)$$

and observe that for the permutation $x_1x_2x_3 \ldots x_n$ we have

$$R = \prod_{1 \leqq i < j \leqq n} (x_j - x_i).$$

This result may also be written in the form

$$R = \prod_{i=1}^{n-1} \prod_{j=2}^{n} (x_j - x_i).$$

5–5 Transpositions. A transposition (ab) has been defined as the interchange of any two elements a and b in a permutation. In this section we shall be primarily concerned with transpositions of adjacent elements. Given the permutation 4132, we may use the sequence of transpositions of adjacent elements

(5–6) (14), (23), (24), (34)

to obtain the sequence of permutations

(5–7) 4132, 1432, 1423, 1243, 1234,

starting with the given permutation and terminating with the elements in their natural order. Although this may be done in several ways, we have for convenience simply considered the numbers 1, 2, 3, 4 in order and successively obtained each in its proper place in the permutation. In general, given any permutation of the elements $a_1, a_2, a_3, \ldots, a_n$, say

(5–8) $$a_{j_1}a_{j_2}a_{j_3} \ldots a_{j_n},$$

the element a_1 must be present as some a_{j_k}. If $a_1 = a_{j_2}$, a single transposition $(a_{j_1}a_1)$ will place a_1 in its proper place (relative to the natural order of the elements). If $a_1 = a_{j_3}$, two transpositions of adjacent elements $(a_{j_2}a_1)$ and $(a_{j_1}a_1)$ may be used. In general, if $a_1 = a_{j_k}$, then $k-1$ transpositions of adjacent elements may be used. Similarly, after obtaining a_1 as the first element, we may consider the new permutation in the form

$$a_1a_{r_1}a_{r_2} \ldots a_{r_{n-1}},$$

where $a_2 = a_{r_s}$ and $s-1$ transpositions of adjacent elements may be used to obtain the permutation

$$a_1a_2a_{t_1}a_{t_2} \ldots a_{t_{n-2}}.$$

Since the permutation (5–8) contains only a finite number of elements, this process may be continued until all the elements are in their natural order. We shall be primarily concerned with permutations of a finite number of elements, i.e., *finite permutations*, and have proved

Theorem 5–1. *The elements of any finite permutation may be obtained in their natural order by a finite sequence of transpositions of adjacent elements.*

The sequence of permutations (5–7) arose when we used the sequence of transpositions (5–6) to obtain the elements of the permutation 4132 in their natural order. Let us now consider the related problem of obtaining the permutation 4132 from the natural order

of its elements 1234. If we start with the permutation 1234, the interchange of 1 and 4 as indicated in (5–6) does not represent a transposition of adjacent elements in the permutation. However, if we take the sequence of transpositions

$$(34),\quad (24),\quad (23),\quad (14)$$

obtained by considering the sequence of transpositions (5–6) in its reverse order, we have the permutations

$$1234,\quad 1243,\quad 1423,\quad 1432,\quad 4132,$$

i.e., the sequence (5–7) in reverse order. In general, we have

THEOREM 5–2. *If an ordered sequence S of transpositions of adjacent elements may be used upon a given permutation to obtain the elements of that permutation in their natural order, then the sequence of transpositions obtained by taking the transpositions of the sequence S in their reverse order may be used upon the permutation of the elements in their natural order to obtain the given permutation.*

The proof of this theorem is an immediate consequence (Exercise 3) of the assumed sequence S, the corresponding sequence of permutations, and the fact that the transpositions (ab) and (ba) have the same effect upon the permutation.

Given any permutation (5–8) let us now consider the effect of a transposition of adjacent elements $(a_{j_k}a_{j_{k+1}})$ upon the class (even or odd) of the given permutation. For any $r < k$ or $r > k+1$ the order of a_{j_r} and a_{j_k} as well as the order of a_{j_r} and $a_{j_{k+1}}$ is not affected by the transposition. Thus the only effect upon the class of the permutation arises from the replacement of $a_{j_k}a_{j_{k+1}}$ by $a_{j_{k+1}}a_{j_k}$. This replacement introduces a new inversion if $a_{i_k}a_{i_{k+1}}$ was a permanence, removes an inversion if $a_{j_k}a_{j_{k+1}}$ was an inversion. Thus a single transposition of adjacent elements always changes the number of inversions by one and we have

THEOREM 5–3. *The class of any permutation is changed by a single transposition of two of its adjacent elements.*

The above three theorems may be used in several of the following exercises to indicate relationships between the class of a permutation and certain sequences of transpositions of its elements. In the next section we shall find that very similar relationships hold when the elements interchanged are not necessarily adjacent in the permutation.

EXERCISES

1. Give a sequence of transpositions of adjacent elements for each of the following permutations that may be used to place the elements of the permutation in their natural order: 3214, *adcb*, 152634, *ptqsr*.

2. Repeat Exercise 1 for the permutations 152634 and *ptqsr* in several ways.

3. Prove Theorem 5–2.

4. Give at least three different sequences of transpositions of adjacent elements that may be used to obtain the permutation 4132 from 1234.

5. Give a sequence of transpositions of adjacent elements for each of the permutations in Exercise 1 that may be used to obtain the permutation from the natural order of its elements.

6. Repeat Exercise 5 for the permutations 152634 and *ptqsr* in several ways.

7. Prove that for any positive integer n we may obtain any given permutation of n elements from the permutation of its elements in their natural order by a sequence of at most $n(n-1)/2$ transpositions of adjacent elements.

8. Verify that each of the odd permutations in Exercises 1, 2, 4, 5, 6 has been changed into or obtained from the natural order of its elements by an odd number of transpositions. Repeat this exercise for the even permutations.

9. Prove that all permutations of $a_1, a_2, \ldots, a_n$ may be obtained using only transpositions of the form $(a_j a_n)$, where n is fixed and j may take on the values $1, 2, \ldots, n-1$.

10. Is it ever possible to obtain an even permutation from the natural order of its elements by an odd number of transpositions of adjacent elements? Explain.

11. Prove that for n greater than 1 exactly half of the $n!$ permutations of n elements are even.

12. Use R as in Section 5–4, Exercise 4, to give a second proof of Theorem 5–3.

5–6 Even and odd permutations. We have seen (Section 5–4) how to count the number of inversions in any given permutation and classify the permutation as even or odd according as the number of inversions is even or odd. We have also seen (Section 5–5) that any given permutation may be obtained from or transformed into a permutation of the elements in natural order by a sequence of transpositions of adjacent elements. Since every such transposition has been associated (Theorem 5–3) with a single inversion in this process, every even permutation can be obtained from or transformed

into the natural order of its elements by an even number of transpositions of adjacent elements (Theorems 5–1 and 5–2). Similarly, every odd permutation can be obtained from or transformed into the natural order of its elements by an odd number of transpositions of adjacent elements. We now prove that any transposition, i.e., any interchange of two elements (adjacent or not) of a permutation, may be obtained by an odd number of transpositions of adjacent elements. Accordingly, we shall prove that any transposition of the elements of a permutation changes the class of the permutation.

Given any permutation, we know (Theorem 5–3) that the interchange of two adjacent elements changes the class of the permutation. The interchange of two elements having a single element between them may be accomplished by three transpositions of adjacent elements. For example, the transpositions

$$(ab) \quad (ac) \quad (bc)$$

may be used to interchange the elements a and c in the permutation abc. The corresponding sequence of permutations is

$$abc, \quad bac, \quad bca, \quad cba.$$

The interchange of two elements having two elements between them in a permutation may be accomplished by five transpositions of adjacent elements. For example, a and d in $abcd$ may be interchanged using the sequence of transpositions

$$(ab) \quad (ac) \quad (ad) \quad (cd) \quad (bd),$$

giving the sequence of permutations

$$abcd, \quad bacd, \quad bcad, \quad bcda, \quad bdca, \quad dbca.$$

In general, the interchange of two elements having k elements between them in a permutation may be accomplished using $2k + 1$ transpositions of adjacent elements (Exercise 1). Thus from Theorem 5–3 the class of any permutation is changed by a single interchange of any two of its elements. In other words, we have proved

THEOREM 5–4. *The class of any permutation is changed by any transposition of its elements.*

The product of the differences as in Exercise 4, Section 5–4, may be used (Exercise 2) to give a second proof of Theorem 5–4. This theorem will be needed in Section 5–8 in proving one of the basic properties of determinants. We now turn from our discussion of properties of permutations to the use of permutations in the definition of the determinant of a square matrix.

EXERCISES

1. Prove that in a permutation the interchange of two elements having k elements between them may be accomplished using $2k + 1$ transpositions of adjacent elements.

2. Use the method of Exercise 12, Section 5–5, to give a second proof of Theorem 5–4.

3. Prove that any given permutation of n elements may be obtained from the natural order of the elements by a sequence of at most $n - 1$ transpositions.

4. Replace "transposition of adjacent elements" by "transposition" in Theorem 5–2 and prove the resulting theorem.

5. Consider arbitrary transpositions of elements (not necessarily adjacent) of the permutation 123 and prove that the resulting permutation depends upon the order in which the transpositions are applied, i.e., that the application of a sequence of transpositions is not necessarily a commutative operation.

6. Use the sequences of transpositions (21), (24); (43), (42), (41), (23); (24), (14) and the natural order 1234 to show that the permutation 4132 may be obtained from the natural order using several different sequences of transpositions.

7. Give at least four different sequences of transpositions that may be used to obtain the permutation 1234 from 4132.

8. Give several examples of sequences of transpositions that are (a) commutative, (b) noncommutative.

5–7 Determinants. In this section we shall consider one explicit procedure for writing down the determinant of any given square matrix. As mentioned in Section 5–2, the determinant of a square matrix is defined to be a polynomial in the elements of the matrix. This polynomial may be obtained in several ways. We shall be primarily interested in two of these methods: the row expansion of a determinant of a square matrix and the column expansion of a determinant of a square matrix. These expansions differ only in the method used to obtain the terms of the polynomial. They will be proved to be equivalent in Section 5–8.

Formally, the *row expansion of the determinant* of a square matrix may be defined as the algebraic sum of all possible products obtainable by taking one and only one factor from each row and column of the

matrix, where each product is preceded by a plus sign or a minus sign according as the number of inversions of the column indices of the factors are even or odd when the row indices are in their natural order [16; 3]. Let us consider a few examples of this definition.

Given any square matrix with two rows and two columns

$$\begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \tag{5–9}$$

we may designate the determinant of this matrix by

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \tag{5–10}$$

and seek the row expansion of the determinant. Thus (5–10) represents the determinant of (5–9) and (5–9) is the matrix of (5–10). By definition, the determinant of (5–9) is the polynomial $a_{11}a_{22} - a_{12}a_{21}$.

According to the above formal definition of the row expansion of a determinant, we may select any element, say a_{11}, and take with it an element that is not on the same row or column as a_{11} in the matrix of the determinant. In other words, if we select a_{11}, we delete the first row and the first column

$$\begin{vmatrix} \cancel{a_{11}} & \cancel{a_{12}} \\ \cancel{a_{21}} & a_{22} \end{vmatrix}$$

and choose an element from those remaining. In the case of (5–10) there remains only one element, and we have the product $a_{11}a_{22}$. In general, we repeat the process by deleting the row and column of the newly chosen element until only a single element remains. Having selected the product $a_{11}a_{22}$ from (5–9) or (5–10), there remains only one other product $a_{12}a_{21}$. Each of these products may be written in two ways: $a_{11}a_{22}$ or $a_{22}a_{11}$ and $a_{12}a_{21}$ or $a_{21}a_{12}$. According to the above definition, we shall take them in the forms $a_{11}a_{22}$ and $a_{12}a_{21}$ in which the row indices (first subscripts) are in their natural order. Then we take each product with a plus sign if the column indices (second subscripts) form an even permutation, a minus sign if the column indices form an odd permutation. Thus the determinant of (5–9) may be expressed as

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = a_{11}a_{22} - a_{12}a_{21}. \tag{5–11}$$

If we had started with a different element, say a_{21} instead of a_{11}, we would have obtained an equivalent expression, such as $-a_{12}a_{21} + a_{11}a_{22}$, for the row expansion of the determinant.

Before considering other examples of the above definition, let us define the order of a square matrix. The number of rows (columns) in a square matrix is called its *order*. Thus (5–9) is a matrix of order 2 and (5–2) is a matrix of order n. Similarly, (5–10) represents a determinant of order 2 and, in general, the order of the determinant of a square matrix is the same as the order of the matrix.

We have applied the above definition of the row expansion of a determinant to determinants of order 2. A determinant of order 3 may be similarly expanded (Exercise 1) in terms of $3! = 6$ products of three factors each as follows:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} = a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31}. \tag{5–12}$$

The polynomial in (5–11) may be expressed in the form

$$\sum e_{j_1 j_2} a_{1 j_1} a_{2 j_2},$$

where $\sum$ is the summation symbol, the summation is taken over the $2! = 2$ permutations of the second subscripts, and $e_{j_1 j_2}$ is taken as $+1$ or -1 according as the permutation of the second subscripts is even or odd relative to the natural order of the positive integers. Similarly, the polynomial in (5–12) may be expressed in the form

$$\sum e_{j_1 j_2 j_3} a_{1 j_1} a_{2 j_2} a_{3 j_3},$$

where the summation is taken over the $3! = 6$ permutations of the second subscripts. In general, the row expansion of the determinant (5–2) of a matrix of order n may be expressed in the form

$$\sum e_{j_1 j_2 \ldots j_n} a_{1 j_1} a_{2 j_2} \ldots a_{n j_n}, \tag{5–13}$$

where the summation is taken over the $n!$ permutations of the second subscripts and the e's are, as before, $+1$ or -1 according as the permutations of the second subscripts are even or odd. Since this general expansion of a determinant of order n is a polynomial involving only ring operations on the elements of the determinant, we may expect to find significance for determinants and matrices with elements from any integral domain (see introductory paragraphs of Chapter 2). The most common matrices and determinants in elementary mathematics have arbitrary real numbers as elements. We shall assume that such is the case throughout most of this chapter, but we shall also consider some applications of matrices with poly-

nomials as elements. Matrices with complex numbers as elements play an important role in advanced mathematical theories.

In the next section we shall consider three properties of determinants that, especially in the case of $n > 3$, will often enable us to simplify the above formal method for expanding a given determinant. For $n = 2$ the polynomial (5–11) is easily obtained as the product of the elements on the principal diagonal (equal indices) diminished by the product of the elements on the minor diagonal. For $n = 3$ there also exists a similar method using diagonal lines. Since many readers have previously used the method in (5–14), we mention it in order to emphasize that there does not exist a similar method when n is greater than 3. For $n = 3$ one may recopy the first two columns as follows:

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \begin{matrix} a_{11} & a_{12} \\ a_{21} & a_{22} \\ a_{31} & a_{32} \end{matrix}, \tag{5–14}$$

add the products of the elements on diagonals parallel to the principal diagonal

$$a_{11}a_{22}a_{33} + a_{12}a_{23}a_{31} + a_{13}a_{21}a_{32},$$

and from this subtract the sum of the products of elements on diagonals parallel to the minor diagonal, i.e., add the negatives

$$-a_{12}a_{21}a_{33} - a_{11}a_{23}a_{32} - a_{13}a_{22}a_{31}$$

of these products. However, for determinants of order 4 this method would give only $2n = 8$ terms of the polynomial, whereas $n! = 24$ terms are needed. In general, the above method of diagonals (5–14) may be used only for determinants of order less than or equal to 3. Other convenient methods (Sections 5–9 and 5–10) may be used for determinants of all orders and are recommended in place of that in (5–14) for determinants of order 3.

Exercises

1. Use the formal definition of the row expansion of a determinant to obtain (5–12).

2. Find the row expansion of

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \end{vmatrix} \quad \text{and} \quad \begin{vmatrix} 2 & 3 \\ 4 & 6 \end{vmatrix}.$$

3. Find the row expansion of

$$\begin{vmatrix} 1 & 2 & 3 \\ 1 & 0 & 1 \\ 2 & 1 & 2 \end{vmatrix}.$$

4. Find the row expansion of

$$\begin{vmatrix} 2 & 1 & 5 \\ 3 & -2 & 2 \\ 1 & -1 & 0 \end{vmatrix}.$$

5. Use (5–2) with $n = 4$ and write down a general matrix of order 4. Find the row expansion of the determinant of this matrix.

6. Find and simplify the row expansion of

$$\begin{vmatrix} 1 & 2 & 3 & 1 \\ 2 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{vmatrix}.$$

7. Define the column expansion of the determinant of a square matrix by interchanging the words "row" and "column" in the definition of the row expansion of a determinant.

8. Repeat Exercise 2, using the column expansion.

9. Find the column expansion of the determinant of (5–9) and compare with the row expansion.

10. Find the column expansion of (5–12) and compare with the row expansion.

11. Repeat Exercises 3 and 4, using the column expansion.

12. Find the column expansion of a general matrix of order 4 and compare with the row expansion obtained in Exercise 5.

13. Repeat Exercise 6, using the column expansion.

14. Use the summation notation as in (5–13) to express the column expansions of square matrices of order 2, 3, 4, and n.

5–8 Properties of determinants. In this section we shall use the row expansions

$$a_{11}a_{22} - a_{12}a_{21}, \tag{5–15}$$

$$a_{11}a_{22}a_{33} - a_{11}a_{23}a_{32} + a_{12}a_{23}a_{31} - a_{12}a_{21}a_{33} + a_{13}a_{21}a_{32} - a_{13}a_{22}a_{31} \tag{5–16}$$

for the determinants (5–11) and (5–12) of general second and third order matrices to illustrate three basic properties of determinants. We shall use the word *line* of a matrix to indicate a row or column of the matrix. This notation will be very useful when a statement applies equally to rows and columns.

Each term of the polynomial (5–15) has exactly one factor with first subscript 1, that is, each term has exactly one factor from the first row of the matrix of the determinant. Similarly, each term has exactly one factor from the second row, first column, second column. Thus each term of the row expansion (5–15) has exactly one factor

from each line of the matrix of the determinant (5–11). In other words, the row expansion (5–15) is linear and homogeneous in the elements of each line of the matrix of the determinant.

The above statements may also be made with reference to the row expansion (5–16) of the determinant (5–12). Each term of the row expansion contains exactly one factor from each line of the matrix of the determinant, i.e., the row expansion is linear and homogeneous in the elements of each line of the matrix of the determinant. When expressed for determinants of any order n, these statements illustrate our first basic property of determinants.

PROPERTY A. *The row expansion of a determinant is linear and homogeneous in the elements of each line of its matrix.*

We next use this property and consider a few particular methods of expanding the general third order determinant (5–12). If in (5–16) we collect the coefficients of the elements of the first row of the determinant, we have

$$(5\text{–}17)\quad a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}),$$

where the term containing a_{1j} has been added when $1 + j$ is even and subtracted when $1 + j$ is odd. The significance of this convention will soon be appreciated when minors of elements are discussed.

The expansion (5–17) is called an expansion in terms of the elements of the first row of the matrix of the determinant. Similarly, the general third order determinant may be expanded in terms of the elements of any line of the matrix of the determinant (Exercise 1).

If the elements of the first row of the matrix of (5–12) are all zero, that is, $a_{11} = a_{12} = a_{13} = 0$, then the expansion (5–17) is zero. Since any determinant may be expressed in terms of the elements of any line of the matrix of the determinant using Property A, we have

THEOREM 5–5. *If all the elements of a line of a square matrix are zero, its determinant is zero.*

Returning to (5–17) we see that the coefficient of a_{11} is precisely the expansion of the determinant obtained by deleting from (5–12) the row and column of a_{11}. The same is true for a_{13} and, except for sign, for a_{12}. In general, we define the *minor of an element* of a square matrix of order n to be the determinant of order $n - 1$ obtained by deleting the row and column in which the element is situated. Then we define the *cofactor* A_{ij} of an element a_{ij} to be the coefficient

of a_{ij} in the determinant (5–13). We have observed that the cofactors of a_{11} and a_{13} in (5–17) are equal to their minors, whereas the cofactor of a_{12} is the negative of the minor of that element. By Property C and (5–13) in Exercise 15 of this section and by another method in Exercise 4, Section 5–10, we may prove that the cofactor of any element a_{ij} is $(-1)^{i+j}$ times the minor of a_{ij}.

Using the above definition of cofactor, the expansion of a determinant of order n may be expressed as

$$a_{11}A_{11} + a_{12}A_{12} + \cdots + a_{1n}A_{1n}$$

in terms of the minors of the elements of the first row of the matrix of the determinant, or as

$$(5\text{–}18)\qquad a_{i1}A_{i1} + a_{i2}A_{i2} + \cdots + a_{in}A_{in}$$

in terms of the minors of the elements of the ith row, or as

$$(5\text{–}19)\qquad a_{1j}A_{1j} + a_{2j}A_{2j} + \cdots + a_{nj}A_{nj}$$

in terms of the minors of the elements of the jth column. It is customary to speak as above of an expansion of a determinant in terms of the minors of the elements instead of using the word cofactor for the minors taken with the proper sign. Theorem 5–5 could have been delayed and introduced at this time as an immediate consequence of these expansions.

If each element of the first row of the matrix of (5–12) is multiplied by k, then a_{11}, a_{12}, a_{13} in (5–17) are replaced by ka_{11}, ka_{12}, ka_{13} and the determinant of the matrix has been multiplied by k. Similarly, from (5–18) and (5–19), we have

THEOREM 5–6. *If the elements of any line of a matrix are multiplied by k, then its determinant is multiplied by k.*

This theorem gives us the right to remove a common factor from any line of the matrix of a determinant and multiply it by the determinant of the new matrix. For example, we have

$$\begin{vmatrix} 4 & 6 \\ 1 & 3 \end{vmatrix} = 2\begin{vmatrix} 2 & 3 \\ 1 & 3 \end{vmatrix} = 6\begin{vmatrix} 2 & 1 \\ 1 & 1 \end{vmatrix} = 6.$$

Theorem 5–6 may also be proved directly from the expansion (5–13) and Property A (Exercise 8).

The two remaining basic properties of determinants may be

illustrated respectively by interchanging the rows and columns of the matrix of a determinant

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \end{vmatrix}$$

and by interchanging two rows of the matrix of a determinant and changing the sign of the determinant:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} = - \begin{vmatrix} a_{21} & a_{22} \\ a_{11} & a_{12} \end{vmatrix}.$$

These properties will be used as a basis for procedures for simplifying the expansion of a determinant (Section 5–9). The term identity as used in the following statement is defined in Section 3–4.

PROPERTY B. *The row expansion of the determinant of a square matrix and the column expansion of the determinant are identical.*

PROPERTY C. *The interchange of any two parallel lines of a square matrix changes the sign of its determinant.*

Both of these properties may be proved from the expansion (5–13). The term "two parallel lines" is used to refer to two rows or two columns. Property B states that

$$\sum e_{i_1 i_2 \cdots i_n} a_{1 i_1} a_{2 i_2} \cdots a_{n j_n} = \sum e_{k_1 k_2 \cdots k_n} a_{k_1 1} a_{k_2 2} \cdots a_{k_n n},$$

where the summations are taken over the $n!$ permutations of the column and row indices respectively. The two summations each contain all $n!$ possible products of the elements a_{ij} such that no two elements are from the same row and no two are from the same column. Thus it only remains to show that each product (term of the summation) has the same sign when its factors are arranged according to their row indices as when they are arranged according to their column indices. For example, when $n = 3$, we have a term $a_{12}a_{23}a_{31}$. When the row indices are taken in order, the permutation of the column indices is 231, which may be obtained from the natural order by the sequence of transpositions (12), (13) and is therefore even. When the column indices are taken in order, we have $a_{31}a_{12}a_{23}$ and the permutation of the row indices is 312, which may be obtained from the natural order using (23), (13) and is therefore also even. Thus the term $a_{12}a_{23}a_{31}$ in the expansion of a third order determinant is taken positive whether the determinant is expanded by rows or by columns. Basically (see Theorem 5–2), this is true because the same

sequence (12), (13) of transpositions used to obtain the permutation 231 of the column indices from their natural order may also be used in the reverse order (13), (12) on the permutation 231 to put the column indices in their natural order and thereby give $a_{31}a_{12}a_{23}$. In general, the sequence of transpositions used to obtain the permutation of the column indices of $a_{1 i_1} a_{2 i_2} \cdots a_{n i_n}$ from their natural order may be used in the reverse order to arrange the factors according to their column indices $a_{k_1 1} a_{k_2 2} \cdots a_{k_n n}$. Since the same number of transpositions are involved in both sequences of transpositions, the two permutations are either both even or both odd, and each term of the expansion of the determinant has the same sign whether the determinant is expanded by rows or by columns. This completes the proof of Property B and gives a justification for considering the expansions (5–18) and (5–19) equivalent.

Property C may be proved quickly from Theorem 5–4 as follows. If two columns of a matrix are interchanged, every permutation in (5–13) changes its class and every term in the expansion changes its sign. Property B and an expansion by columns may be used to obtain the same result when any two rows are interchanged.

The following theorem is an immediate consequence (Exercise 12) of Property C and Theorem 5–6.

THEOREM 5–7. *A square matrix in which two parallel lines are proportional has determinant zero.*

In the next two sections we shall use the above properties and theorems to develop methods for simplifying the task of expanding the determinant of any given square matrix.

EXERCISES

1. Use (5–16) and give expansions similar to (5–17) of (5–12) in terms of the elements on its (a) second row, (b) third row, (c) first column, (d) third column.
2. Give a square matrix of order 3 illustrating Theorem 5–5.
3. Give a square matrix of order 2 with determinant zero and no zero elements.
4. Give the minor of each element of (5–12).
5. Give the cofactor of each element of (5–12).
6. Repeat Exercise 1, using cofactors.
7. Give the cofactors of each element of the determinant of a general matrix of order 4.

8. Prove Theorem 5–6 directly from (5–13) and Property A.

9. Use Theorem 5–6 and write the following determinants as products of fractions and determinants with integral elements:

$$\begin{vmatrix} \frac{1}{2} & \frac{1}{3} \\ \frac{1}{4} & \frac{1}{6} \end{vmatrix}, \quad \begin{vmatrix} 1 & \frac{1}{2} & \frac{1}{4} \\ 2 & 3 & 5 \\ 2 & \frac{2}{3} & -1 \end{vmatrix}.$$

10. Prove that the determinant of any square matrix having the elements of one line respectively proportional to the corresponding elements of a parallel line may be expressed as a constant multiple of the determinant of a matrix having two parallel lines identical.

11. Prove that a matrix having two identical parallel lines has determinant zero.

12. Prove Theorem 5–7.

13. Prove that the cofactor of a_{11} is equal to its minor.

14. Prove that the cofactor of a_{2j} is $(-1)^{2+j}$ times its minor.

15. Prove that the cofactor of a_{ij} is $(-1)^{i+j}$ times its minor.

16. Rewrite the determinant designated by

$$\begin{vmatrix} 3 & 1 & 2 & 5 \\ 4 & -1 & 3 & 1 \\ 6 & 2 & 4 & 3 \\ -1 & 1 & 1 & 2 \end{vmatrix}$$

as a determinant having integral elements and having each element in the first column $+1$.

17. Prove that

$$\begin{vmatrix} yz & 1 & x \\ zx & 1 & y \\ xy & 1 & z \end{vmatrix} = \begin{vmatrix} 1 & x & x^2 \\ 1 & y & y^2 \\ 1 & z & z^2 \end{vmatrix}.$$

18. Prove that $a_{1j}A_{1k} + a_{2j}A_{2k} + \cdots + a_{nj}A_{nk} = 0$ and $a_{j1}A_{k1} + a_{j2}A_{k2} + \cdots + a_{jn}A_{kn} = 0$ when $j \neq k$.

5–9 Expansion of determinants. We have formally defined the row expansion of a determinant (Section 5–7) and considered its basic properties (Section 5–8). The column expansion of a determinant has been defined (Exercise 7, Section 5–7) and proved to be identical with the row expansion (Property C, Section 5–8). The expansions of a determinant of a square matrix in terms of minors of the elements of a given row (5–18) or a given column (5–19) are also identical with the row expansion of the determinant. Thus we may speak of *the* expansion of a determinant and seek ways of reducing the labor of expanding a determinant, i.e., finding the polynomial

associated with any given square matrix. We shall frequently designate determinants by arrays such as (5–20) in order to visualize the rows and columns of the matrix of the determinant.

The row expansion of a determinant of order n has $n!$ terms, whereas the row expansion of a determinant of order $n-1$ has only $(n-1)!$ terms. We shall therefore consider methods of replacing a determinant of order n by a determinant of order $n-1$, changing only the form of the expansion and not its value, i.e., the expansion of the new $(n-1)$st order determinant must be identically equal to that of the given nth order determinant. For example, if all but one of the elements on a line of the matrix of a determinant of order n are zero, this determinant may be replaced by a determinant of order $n-1$ using one of the expansions (5–18), (5–19). When $n = 3$, we have the relations

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ 0 & a_{22} & a_{23} \\ 0 & a_{32} & a_{33} \end{vmatrix} = a_{11}\begin{vmatrix} a_{22} & a_{23} \\ a_{32} & a_{33} \end{vmatrix} = \begin{vmatrix} a_{11} & 0 & 0 \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}$$

for arbitrary elements a_{ij}. In this section we shall consider two theorems that will enable us to use the above procedures for the determinant of any square matrix.

The determinant designated by

$$\begin{vmatrix} a_{11}+b_{11} & a_{12}+b_{12} & a_{13}+b_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} \tag{5–20}$$

may be expanded (5–17) in terms of the elements of the first row of its matrix as follows:

$$(a_{11}+b_{11})(a_{22}a_{33} - a_{23}a_{32}) - (a_{12}+b_{12})(a_{21}a_{33} - a_{23}a_{31}) + (a_{13}+b_{13})(a_{21}a_{32} - a_{22}a_{31}).$$

This expansion may be rewritten in the form

$$a_{11}(a_{22}a_{33} - a_{23}a_{32}) - a_{12}(a_{21}a_{33} - a_{23}a_{31}) + a_{13}(a_{21}a_{32} - a_{22}a_{31}) + b_{11}(a_{22}a_{33} - a_{23}a_{32}) - b_{12}(a_{21}a_{33} - a_{23}a_{31}) + b_{13}(a_{21}a_{32} - a_{22}a_{31}),$$

which, when compared with (5–17), is seen to represent the sum of two determinants. This sum may be designated by

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} b_{11} & b_{12} & b_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix}. \tag{5–21}$$

In general, we may use the above procedure and prove (Exercise 1)

THEOREM 5–8. *If the jth row (or column) of a matrix M consists of elements of the form $a_{ji} + b_{ji}$, then the determinant D of M satisfies $D = D_1 + D_2$, where D_1 and D_2 are determinants of matrices whose elements are the same as those of M except that the jth rows (columns) are respectively the a_{ji} and the b_{ji}.*

When $b_{ji} = ka_{ti}$, $t \neq j$, Theorem 5–8 has a very useful application. For example, if $b_{1i} = ka_{2i}$ in (5–20), then (5–21) becomes

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix} + \begin{vmatrix} ka_{21} & ka_{22} & ka_{23} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{vmatrix},$$

where the second determinant is zero, by Theorem 5–7. Similarly any third order determinant is unchanged if a fixed constant multiple (positive or negative) of the corresponding element on the third row of its matrix is added to each element of the first row. We may also prove that a fixed constant multiple of the elements of any row of the matrix of a third order determinant may be added to the corresponding elements of any other row without changing the determinant. A similar statement may be made for the columns of the matrix of a third order determinant (Exercise 4) and, in general, (Exercise 5) for rows and columns of the matrix of any determinant using (5–18) and (5–19). Thus we have

THEOREM 5–9. *The determinant of any square matrix is unchanged by adding to the elements of any line of the matrix a fixed constant multiple of the corresponding elements of any distinct parallel line.*

There are two precautions to be observed in the application of Theorem 5–9. First, one cannot add k times the elements of a line to the elements of the same line, since this would multiply the determinant by $k + 1$ (Theorem 5–6). Second, the new elements, say $a_{ij} + ka_{tj}$, must replace the elements a_{ij}. If they were used to replace the elements a_{tj}, the determinant would be, in effect, multiplied by k. With these precautions, Theorem 5–9 is extremely useful in rewriting a determinant so that at most one of the elements on some line of its matrix is different from zero. Then (5–18) or (5–19) may be used to express the given determinant as a constant times a determinant of lower order. If $a_{11} = 1$, then a_{21} times each element of the first row of the matrix of the determinant may be subtracted from the corresponding element of the second row, and thus a_{21} may be replaced by zero. Similarly, except for a_{11}, each element of the

first row and each element of the first column may be replaced by zero (Exercise 7). If some a_{ij} satisfies $|a_{ij}| = 1$ in any determinant, then every other element of the ith row and every other element of the jth column may be replaced by zero. In general, the determinant of any square matrix with arbitrary real numbers as elements may be expressed as the determinant of a matrix such that at most one element on each row and column is different from zero (Exercise 16).

The determinant designated by

$$\begin{vmatrix} 2 & 5 & 7 \\ 4 & 3 & 2 \\ 2 & 4 & 5 \end{vmatrix} \tag{5–22}$$

may be expanded, using the above principles, as follows: the common factor 2 may be removed from the first column of the matrix of the determinant and the determinant of the new matrix multiplied by 2 (Theorem 5–6), the second row may be decreased by twice the third row (Theorem 5–9), and the first row may be decreased by the third. Taking these steps in the order stated, we have the following:

$$2\begin{vmatrix} 1 & 5 & 7 \\ 2 & 3 & 2 \\ 1 & 4 & 5 \end{vmatrix} = 2\begin{vmatrix} 1 & 5 & 7 \\ 0 & -5 & -8 \\ 1 & 4 & 5 \end{vmatrix} = 2\begin{vmatrix} 0 & 1 & 2 \\ 0 & -5 & -8 \\ 1 & 4 & 5 \end{vmatrix}.$$

We next expand the determinant designated last in terms of the minors of the elements of the first column of its matrix, as in (5–19), and obtain

$$2 \cdot 1\begin{vmatrix} 1 & 2 \\ -5 & -8 \end{vmatrix} = 2(-8 + 10) = 4.$$

We have employed the usual terminology "expansion of a determinant" to designate the process of obtaining the polynomial or number (i.e., the determinant) associated with any given square matrix. When the elements of the matrix are numbers, the expansion of the determinant is often referred to as an *evaluation* of the determinant. In this sense we have evaluated the determinant designated by (5–22) and found it to have value 4, that is, the determinant is the polynomial 4.

Given any set of elements $b_1, b_2, \ldots, b_k$, we may define

$$c_1b_1 + c_2b_2 + \cdots + c_kb_k,$$

where the c's are constants and at least one $c_j \neq 0$ to be a *linear combination* of the b's. Then Theorem 5–9 may be extended (Exercise 20) to read: The determinant of any square matrix is unchanged

by adding to the elements of any line of the matrix any fixed linear combination of the corresponding elements of the remaining parallel lines. An example and further development of this concept will be given in connection with linear dependence (Section 5–13).

We may now expand determinants of square matrices of order n for any positive integer n using minors of the elements of any line. Theorem 5–9 may be used to reduce the number of terms in the expansion. It is thus frequently advantageous to expand a determinant of order n in terms of determinants of order $k < n$ having elements from the given determinant. For $k = n - 1$, such expansions are given by (5–18) and (5–19). For $k < n - 1$, we shall consider new methods in Section 5–10.

Exercises

1. Prove Theorem 5–8.

2. Give and check an example of Theorem 5–8, using determinants of order 2 with numerical elements.

3. Repeat Exercise 2 for determinants of order 3.

4. Prove that the elements of any column of a third order matrix may be increased or decreased by a fixed constant multiple of the corresponding elements of any other column without changing the determinant of the matrix.

5. Prove Theorem 5–9.

6. Give an example of Theorem 5–9, using a determinant of order 3.

7. Write down a third order determinant having $a_{22} = -1$ and $|a_{ij}| > 1$ when $a_{ij} \neq a_{22}$. Use Theorem 5–9 to rewrite this determinant so that a_{21}, a_{23}, a_{12}, and a_{32} are replaced by zero.

8. Repeat Exercise 7 for a determinant of order 4 in which a_{21}, a_{23}, a_{24}, a_{12}, a_{32}, and a_{42} are to be replaced by zero.

9. Evaluate the determinants

$$\begin{vmatrix} 1 & 2 & 3 \\ 1 & 4 & 3 \\ 1 & 5 & 4 \end{vmatrix}, \quad \begin{vmatrix} 2 & 4 & 6 \\ 11 & 4 & 5 \\ 1 & 2 & 3 \end{vmatrix}, \quad \begin{vmatrix} 2 & 5 & 7 \\ 3 & 5 & 6 \\ 7 & 6 & 5 \end{vmatrix}.$$

[Answers $+2, 0, -6$.]

10. Expand

$$\begin{vmatrix} 1 & 0 & 0 & 0 \\ a & 1 & 0 & 0 \\ b & c & 1 & 0 \\ d & e & f & 1 \end{vmatrix}.$$

11. State and prove a general theorem for the expansion of determinants of triangular matrices such as that in Exercise 10 in which all elements above the principal diagonal are zero.

12. Evaluate

$$\begin{vmatrix} 2 & 7 & 5 & 6 \\ 1 & 6 & 4 & 5 \\ 2 & 3 & 4 & 2 \\ 3 & 2 & 1 & 4 \end{vmatrix}.$$

13. Expand the determinants

$$\begin{vmatrix} x & y & z & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{vmatrix}, \quad \begin{vmatrix} x & y & z & 1 \\ 1 & 1 & 1 & 1 \\ 2 & 3 & 4 & 1 \\ -1 & 2 & 5 & 1 \end{vmatrix}.$$

14. Use the Euclidean Algorithm (Section 2–5) and prove that any second order determinant (5–10) having integral elements may be rewritten such that a_{12} and a_{21} are replaced by zero, i.e., any second order determinant with integral elements may be rewritten so that at most the elements on the principal diagonal are different from zero. (It is sufficient to use integers or, in general, elements of an integral domain).

15. Prove that the third order determinant (5–12) with real numbers as elements may be rewritten so that at most the elements on the principal diagonal are different from zero.

16. Outline a procedure by means of which the determinant of any square matrix with arbitrary real numbers as elements may be expressed as the determinant of a matrix such that at most the elements on the principal diagonal of the new matrix are different from zero. Will the given procedure apply when the elements of the determinant are arbitrary elements of any given integral domain in the complex number system?

17. Evaluate

$$\begin{vmatrix} 1 & 2 & 1 & 3 & 1 \\ 2 & 1 & 3 & 2 & 2 \\ 1 & 1 & 4 & 2 & 1 \\ 3 & 4 & 1 & 2 & 2 \\ 2 & 4 & 2 & 4 & 2 \end{vmatrix}.$$

[Answer $+16$.]

18. Express the following determinant as the sum of two third order determinants:

$$\begin{vmatrix} x + a & y + b & z + c \\ 1 & 2 & 3 \\ 4 & 1 & 2 \end{vmatrix}.$$

19. Express the following sum of two determinants as a single third order determinant:

$$\begin{vmatrix} 5 & -6 & 7 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{vmatrix} + \begin{vmatrix} 10 & 7 & -5 \\ 1 & 2 & 3 \\ 1 & 1 & 1 \end{vmatrix}.$$

20. Prove that to the elements of any line of a square matrix may be added any fixed linear combination of the corresponding elements of the remaining parallel lines without changing the determinant of the matrix.

21. Prove that

$$\begin{vmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ 1 & x_3 & x_3^2 \end{vmatrix} = (x_3 - x_1)(x_2 - x_1)(x_3 - x_2).$$

22. Extend Exercise 21 to show that for any integer k Vandermonde's determinant

$$\begin{vmatrix} 1 & x_1 & x_1^2 & \dots & x_1^{k-1} \\ 1 & x_2 & x_2^2 & \dots & x_2^{k-1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 1 & x_k & x_k^2 & \dots & x_k^{k-1} \end{vmatrix}$$

$$= (x_k - x_1)(x_k - x_2) \dots (x_k - x_{k-1})$$
$$(x_{k-1} - x_1)(x_{k-1} - x_2) \dots (x_{k-1} - x_{k-2})$$
$$\cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot$$
$$(x_3 - x_1)(x_3 - x_2)$$
$$(x_2 - x_1).$$

5–10 Minors. We have defined a square array such as (5–2) to be a square matrix, associated a determinant with each square matrix, and (Section 5–8) defined the minor of an element of a matrix of order n to be the determinant of order $n - 1$ obtained by deleting the row and column on which the element is situated. This definition will now be extended as follows: Given any matrix of order n, the matrix obtained by deleting any r ($r < n$) rows and any r columns of the given matrix has a determinant of order $n - r$ that is called an rth *minor* of the given matrix. Thus the minor of any element a_{ij} of a matrix is a first minor of the matrix. The determinant obtained in (5–23) by deleting the first and second columns, the second and fourth rows is

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} = \begin{vmatrix} a_{13} & a_{14} \\ a_{33} & a_{34} \end{vmatrix} \tag{5–23}$$

and is called a second minor of the given fourth order matrix.

Given a matrix of order n, we may obtain an rth minor either by deleting r rows and r columns or by choosing in their natural order

$n - r$ rows and $n - r$ columns from which the elements are to be taken. For example, the minor of a_{11} in a third order matrix may be obtained by deleting the first row and first column or by selecting the elements that are on the second or third rows and second or third columns. Thus in counting the first minors of a third order matrix, we may say that there is a minor associated with each of the 3^2 elements of the matrix, or we may calculate $C_2^3 = (3 \cdot 2)/2 = 3$ the number of ways of choosing two rows out of three and also the number of ways of choosing two columns out of three whence, as before, there are $(C_2^3)^2 = 9$ first minors of a third order matrix. In general, (Exercise 1) there are $(C_{n-r}^n)^2$ rth minors of an nth order matrix, where $C_{n-r}^n = n!/[(n-r)!\,r!]$ and $n! = 1 \cdot 2 \cdot 3 \cdot \dots \cdot n$.

When the rows and columns used in forming a minor M_2 are precisely those left unused in forming a minor M_1, the two minors M_1 and M_2 are called *complementary minors*. For example,

$$\begin{vmatrix} a_{21} & a_{22} \\ a_{41} & a_{42} \end{vmatrix} \text{ and } \begin{vmatrix} a_{13} & a_{14} \\ a_{33} & a_{34} \end{vmatrix}$$

are complementary minors of the matrix of (5–23). The *algebraic complement* of a minor of a matrix is equal to its complementary minor multiplied by $(-1)^p$, where p is the sum of the indices of the rows and columns used in the formation of the minor (Exercises 3–6 and [9; 23–24]). Since the sum of all row and column indices of any square matrix of order n is an even number $n(n-1)$, we could instead take p as the sum of the indices of the rows and columns deleted. Thus the algebraic complement of a minor corresponds to the cofactor of an element. In particular, since an $(n-1)$st minor is a single element, the algebraic complement of any element of a matrix is its cofactor.

The determinant of any matrix may be obtained by selecting any row of the matrix, multiplying each element on that row by its algebraic complement, and taking the sum of these products, as in (5–18). This procedure may also be used for any column of the matrix, as in (5–19). These expansions in terms of $(n-1)$st minors of all elements on a line of the matrix are special cases of the following theorem.

THEOREM 5–10. LAPLACE'S EXPANSION. *Select any r parallel lines from a matrix M and form all the r-rowed minors from this array. The determinant of M is equal to the sum of the products of each of these minors by its algebraic complement.*

We shall outline the proof of Theorem 5–10, leaving most of the details for the reader to fill in either as an exercise (Exercise 20) or using a more detailed text on matrices and determinants, such as [16; 20–22]. Briefly, one must prove that every term of the determinant occurs exactly once with the proper sign in the Laplace expansion and that no other terms occur. The terms of the row expansion of a determinant (Section 5–7) are defined to be the products obtained by taking one and only one factor from each row and column of the matrix of the determinant. Thus every term of the determinant of a matrix of order n has n factors. We assume that the elements of the matrix are from a ring in which multiplication is commutative. Then the $n!$ terms of the determinant are independent of the permutations of the rows and columns, i.e., it can be shown that each term occurs once and only once whether the expansion is in terms of $(n-1)$st minors or rth minors, where $0 < r < n$. Finally, the method used in Exercise 6 may be extended to prove that the sign of the term is independent of the method of expansion. The details of the proof may be supplied by the reader or found in other texts, such as [16; 20–22].

As mentioned above, the expansions (5–18) and (5–19) are special cases of this theorem in which $r = 1$. The following example illustrates the theorem when $r = 2, n = 4$, and the first two rows are selected.

$$\begin{vmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \cdot \begin{vmatrix} a_{33} & a_{34} \\ a_{43} & a_{44} \end{vmatrix} - \begin{vmatrix} a_{11} & a_{13} \\ a_{21} & a_{23} \end{vmatrix} \cdot \begin{vmatrix} a_{32} & a_{34} \\ a_{42} & a_{44} \end{vmatrix}$$

$$+ \begin{vmatrix} a_{11} & a_{14} \\ a_{21} & a_{24} \end{vmatrix} \cdot \begin{vmatrix} a_{32} & a_{33} \\ a_{42} & a_{43} \end{vmatrix} + \begin{vmatrix} a_{12} & a_{13} \\ a_{22} & a_{23} \end{vmatrix} \cdot \begin{vmatrix} a_{31} & a_{34} \\ a_{41} & a_{44} \end{vmatrix}$$

$$- \begin{vmatrix} a_{12} & a_{14} \\ a_{22} & a_{24} \end{vmatrix} \cdot \begin{vmatrix} a_{31} & a_{33} \\ a_{41} & a_{43} \end{vmatrix} - \begin{vmatrix} a_{13} & a_{14} \\ a_{23} & a_{24} \end{vmatrix} \cdot \begin{vmatrix} a_{31} & a_{32} \\ a_{41} & a_{42} \end{vmatrix}.$$

This procedure is most useful when several of the r-rowed minors of the corresponding matrix are zero, as in (5–24), when the first and second rows are selected and $r = 2$ (Exercise 9).

$$\begin{vmatrix} 1 & 2 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 1 & 2 & 3 & 4 \\ 4 & 3 & 2 & 1 \end{vmatrix}. \tag{5–24}$$

Theorem 5–10 is also important in that it may be used to prove Theorem 5–11 regarding products of square matrices. The same procedure may be used for any two matrices such that the number of columns in the first matrix is equal to the number of rows in the second matrix. The multiplication of matrices is highly important in mathematical theories. We shall consider several applications of this procedure in Section 5–15.

We first define the *inner product* of two ordered n-tuples such as

$$V = a_1, a_2, a_3, \ldots, a_n,$$
$$W = b_1, b_2, b_3, \ldots, b_n$$

to be the sum of the products of corresponding elements,

$$a_1b_1 + a_2b_2 + a_3b_3 + \cdots + a_nb_n.$$

Then we define the *product of the square matrices* $M = [a_{ij}]$ and $N = [b_{ij}]$ of order n to be a square matrix $[c_{ij}]$ of order n, where c_{ij} is the inner product of the set of elements on the ith row of M and the set of elements on the jth column of N, that is,

$$c_{ij} = a_{i1}b_{1j} + a_{i2}b_{2j} + a_{i3}b_{3j} + \cdots + a_{in}b_{nj}.$$

The significance of this definition will be evident when geometric transformations are considered in Section 5–15. It is in part due to the property stated in the following theorem.

THEOREM 5–11. *The determinant of the product of two matrices is equal to the product of their determinants.*

For example, the following equality may be obtained from the above definition and Theorem 5–11:

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix} \cdot \begin{vmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{vmatrix} = \begin{vmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{vmatrix}.$$

This equality may be easily verified by using the polynomials (determinants) designated by the arrays. Also, by Theorem 5–10, the above product is equal to the determinant

$$D = \begin{vmatrix} a_{11} & a_{12} & 0 & 0 \\ a_{21} & a_{22} & 0 & 0 \\ -1 & 0 & b_{11} & b_{12} \\ 0 & -1 & b_{21} & b_{22} \end{vmatrix}.$$

By Theorem 5–9, each element on the third column of the matrix of D may be replaced by itself plus b_{11} times the corresponding

element on the first column plus b_{21} times the corresponding element on the second column. Similarly, the fourth column may be replaced by itself plus b_{12} times the first column plus b_{22} times the second column. Then we have a new matrix with the same determinant D:

$$\begin{vmatrix} a_{11} & a_{12} & a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21} & a_{22} & a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{vmatrix},$$

and by Theorem 5–10 this determinant is equal to

$$\begin{vmatrix} a_{11}b_{11} + a_{12}b_{21} & a_{11}b_{12} + a_{12}b_{22} \\ a_{21}b_{11} + a_{22}b_{21} & a_{21}b_{12} + a_{22}b_{22} \end{vmatrix}.$$

We have used Theorems 5–9 and 5–10 to prove Theorem 5–11 in the special case of two matrices of order two.

In general, given any two square matrices of order n, say $[a_{ij}]$ and $[b_{ij}]$, we may use Theorem 5–10 to write

$$|a_{ij}| \cdot |b_{ij}| = \begin{vmatrix} |a_{ij}| & 0 \\ F & |b_{ij}| \end{vmatrix},$$

where 0 is the determinant of a square matrix of order n with all its elements zeros and F is the determinant of a square matrix of order n with each element on its principal diagonal -1 and all other elements zeros. Then by Theorem 5–9 the $(n+1)$st column of the matrix of this determinant may be replaced by itself plus b_{11} times the first column plus b_{21} times the second column plus . . . plus b_{n1} times the nth column. Similarly, the $(n+1)$st, $(n+2)$nd, . . . and $(2n)$th columns are each replaced by themselves plus multiples of the first n columns. The new matrix has the same determinant as before, by Theorem 5–9, and, as in the above special case, that determinant has the desired form by Theorem 5–10 (Exercise 21).

The importance of Theorems 5–9, 5–10, and 5–11 will be evident in the following exercises and in the remaining sections of this chapter. We have indicated the procedures used in the proofs of these theorems. The detailed proofs have been given as exercises. They may be found in most texts on matrices and determinants.

The concept of a minor of a matrix is used as follows: Given any matrix A of m rows and n columns (5–1), square matrices of order r ($r \leqq m, n$) may be obtained by selecting the elements on any r rows and r columns of the matrix A. Such matrices are called r-rowed minors of the matrix A. The *rank* of the matrix A is the largest

integer r such that A has an r-rowed minor with a determinant different from zero, i.e., there is an r-rowed minor of A with determinant different from zero and every $(r+1)$-rowed minor of A has determinant zero. For example, each of the following matrices has rank two:

$$\begin{bmatrix} 1 & 2 \\ 3 & 5 \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 3 \\ 2 & 2 & 2 \end{bmatrix}.$$

A formal procedure for determining the rank of any matrix is discussed in Exercise 17.

We shall find the concept of the rank of a matrix very useful as we now leave the study of the theory of determinants and matrices and consider some of their applications. Most of the applications will be taken without further introduction from topics usually studied in college algebra and analytic geometry. Rectangular Cartesian coordinate systems will be used throughout the remainder of this text unless otherwise specified.

Exercises

1. Prove that there are $(C_{n-r}^{n})^2$ rth minors of an nth order matrix.
2. List the second minors of a third order matrix.
3. Use (5–18) and the fact that by Theorem 5–8 the determinant of (5–2) may be designated by

$$\begin{vmatrix} a_{11} & 0 & 0 & \dots & 0 \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{vmatrix} + \begin{vmatrix} 0 & a_{12} & a_{13} & \dots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & a_{n3} & \dots & a_{nn} \end{vmatrix}$$

to prove that the algebraic complement (cofactor) of a_{11} is equal to its minor.

4. Use Property C and Exercise 3 to show that the algebraic complement of any element a_{ij} is $(-1)^{i+j}$ times its minor.
5. Extend the results of Exercises 3 and 4 to prove that the algebraic complement of

$$\begin{vmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{vmatrix}$$

is equal to its minor.

6. Prove that the algebraic complement of any second order minor

$$\begin{vmatrix} a_{rs} & a_{ru} \\ a_{ts} & a_{tu} \end{vmatrix}$$

is $(-1)^{r+s+t+u}$ times its minor.

7. Repeat Exercise 10, Section 5–9, using the second order minors on the third and fourth rows of the corresponding matrix.

8. Repeat Exercise 10, Section 5–9, using the second order minors on the third and fourth columns.

9. Expand (5–24) in terms of the second minors on the first two rows of its matrix.

10. A minor obtained by deleting rows and columns of the same indices (for example, first and third rows, first and third columns) is called a *principal minor*. How many second order principal minors are there in a matrix of order n?

11. List the second order principal minors of the matrix of the determinant in Exercise 17, Section 5–9.

12. Repeat Exercise 17, Section 5–9, using the third order minors on the first, second, and third rows of the corresponding matrix.

13. List the 18 pairs of second order minors of the matrix of a general fourth order determinant (5–23).

14. Write down five square matrices of three rows and three columns with numerical elements. Determine the rank of each matrix.

15. Write down a matrix of four rows and five columns and determine its rank.

16. Prove that the rank of a matrix is not affected by any of the following *elementary transformations:* interchange of two parallel lines, multiplication of all elements of a line by a constant different from zero, addition to the elements of one line of multiples of the corresponding elements of another parallel line.

17. A matrix $[a_{jk}]$ is said to be in *normal form* when $a_{jk} = 0$ for $j \neq k$ and, for some integer s, $a_{jj} \neq 0$ for $j \leqq s$, $a_{jj} = 0$ for $j > s$. Use the following matrices to illustrate how any matrix may be replaced by a matrix in the normal form using the elementary transformations:

$$\begin{bmatrix} 1 & 2 & 4 & 6 \\ 2 & 3 & 5 & 7 \\ 11 & 12 & 14 & 16 \\ 5 & 6 & 8 & 10 \end{bmatrix}, \quad \begin{bmatrix} 2 & 3 & 5 & 7 \\ 4 & 2 & 1 & 3 \\ 3 & 6 & 5 & 4 \\ 2 & -2 & 7 & 5 \end{bmatrix}.$$

18. Repeat Exercises 14 and 15, using the method of Exercise 17.

19. Express the following products as single matrices:

(a) $$\begin{bmatrix} a & b & c \\ d & e & f \\ 1 & 1 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ -1 & -1 & 1 \end{bmatrix},$$

(b) $$\begin{bmatrix} x & 0 & 0 \\ a & y & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x & 0 & 0 \\ 0 & y & 0 \\ c & d & e \end{bmatrix} \cdot$$

20. Give a complete proof of Theorem 5–10.

21. Give a complete proof of Theorem 5–11.

5–11 Cramer's Rule. We found in Section 5–1 that the system of equations

$$\begin{aligned} a_1x + b_1y &= c_1, \\ a_2x + b_2y &= c_2 \end{aligned} \tag{5–25}$$

has a unique solution if and only if the determinant

$$D = \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}$$

of the coefficients is different from zero. Furthermore, when $D \neq 0$, the solution is $x = D_1/D$, $y = D_2/D$, where D_1 is obtained by replacing the coefficients of x in D by the constant terms and D_2 is similarly obtained by replacing the coefficients of y by the constant terms, i.e.,

$$D_1 = \begin{vmatrix} c_1 & b_1 \\ c_2 & b_2 \end{vmatrix} \quad \text{and} \quad D_2 = \begin{vmatrix} a_1 & c_1 \\ a_2 & c_2 \end{vmatrix}.$$

A linear equation such as $x + 3y - 5z = 0$, in which the constant term is zero and each term on the left is of the first degree, is called a *linear homogeneous equation.* A linear equation having a constant term different from zero is called a *linear nonhomogeneous equation.* The above method of solving two linear equations in two variables may be extended (Theorem 5–12) to include systems of n linear equations in n variables. It may be stated for arbitrary systems of linear homogeneous equations and for systems of linear nonhomogeneous equations (Section 5–12).

For $n = 3$ we shall show that the system

$$\begin{aligned} a_1x + b_1y + c_1z &= d_1, \\ a_2x + b_2y + c_2z &= d_2, \\ a_3x + b_3y + c_3z &= d_3 \end{aligned} \tag{5–26}$$

has a unique solution $x = D_1/D$, $y = D_2/D$, $z = D_3/D$ if and only if the *determinant of the coefficients*

$$D = \begin{vmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{vmatrix} \neq 0,$$

where D_1 is obtained from D by replacing the a_i's respectively by the constant terms d_i's, D_2 and D_3 are similarly obtained by replacing the b_i's and c_i's by the d_i's. For example, given the system of equations

$$\begin{aligned} x + y + 2z &= 5, \\ x - y + z &= 2, \\ 3x + 2y - 5z &= 7, \end{aligned}$$

we may evaluate the determinant of the coefficients

$$D = \begin{vmatrix} 1 & 1 & 2 \\ 1 & -1 & 1 \\ 3 & 2 & -5 \end{vmatrix} = 21$$

and, since $D \neq 0$, express the unique solution of the system as

$$x = \tfrac{1}{21}\begin{vmatrix} 5 & 1 & 2 \\ 2 & -1 & 1 \\ 7 & 2 & -5 \end{vmatrix} = \tfrac{54}{21} = \tfrac{18}{7},$$

$$y = \tfrac{1}{21}\begin{vmatrix} 1 & 5 & 2 \\ 1 & 2 & 1 \\ 3 & 7 & -5 \end{vmatrix} = \tfrac{25}{21},$$

$$z = \tfrac{1}{21}\begin{vmatrix} 1 & 1 & 5 \\ 1 & -1 & 2 \\ 3 & 2 & 7 \end{vmatrix} = \tfrac{13}{21}.$$

In general, for any system (5–26), we may let A_j be the cofactor (Section 5–8) of a_j in the determinant of the coefficients, multiply both sides of the first equation by A_1, multiply both sides of the second equation by A_2, multiply both sides of the third equation by A_3, and add the resulting equations to obtain $Dx = D_1$, since from (5–19) and Exercise 18, Section 5–8, we have

$$\begin{aligned} a_1A_1 + a_2A_2 + a_3A_3 &= D, \\ b_1A_1 + b_2A_2 + b_3A_3 &= 0, \\ c_1A_1 + c_2A_2 + c_3A_3 &= 0, \end{aligned}$$

and by definition $d_1A_1 + d_2A_2 + d_3A_3 = D_1$. Similarly, we may solve (5–26) for y and z, using cofactors of their coefficients. Finally, for any system of n linear equations in n variables,

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2, \\ &\cdot \\ &\cdot \\ &\cdot \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= b_n, \end{aligned}$$

we may multiply both sides of the jth equation by the cofactor A_{1j} of a_{1j} for $j = 1, 2, \ldots, n$, add the resulting equations, and obtain $Dx_1 = D_1$. Similarly (Exercise 7), we may use the cofactors A_{2j} to obtain $Dx_2 = D_2$, and, in general, A_{kj} to obtain $Dx_k = D_k$. Thus we have

THEOREM 5–12. CRAMER'S RULE. *A system of n linear equations in n variables*

$$a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jn}x_n = b_j, \quad (j = 1, 2, \ldots, n)$$

has a unique solution $x_k = D_k/D$ when $D \neq 0$ is the determinant of the coefficients and D_k $(k = 1, 2, \ldots, n)$ is obtained from D upon replacing the coefficients of x_k by the constant terms.

This theorem may also be considered as giving a sufficient condition that n linear equations in n variables be *consistent*, i.e., have at least one common solution. Using the concept of the rank of a matrix (Section 5–10), we may say that two linear equations in two variables (5–25) are consistent and have a unique common solution if the matrix of the coefficients is of rank two. Similarly, three linear equations in three variables (5–26) are consistent and have a unique common solution if the matrix of the coefficients is of rank three. In general, n linear equations in n variables are consistent and have a unique common solution if the matrix of the coefficients is of rank n. If the rank of the matrix of the coefficients is not equal to n, the system may be inconsistent, as for example,

$$\begin{aligned} x + y &= 1, \\ x + y &= 2, \end{aligned}$$

or the system may be consistent but not have a unique solution. Considered graphically, the two lines

$$\begin{aligned} x + y &= 1, \\ 2x + 2y &= 2 \end{aligned}$$

coincide, the three planes (5–26) might have a line in common or might coincide. Thus the above condition for consistency is sufficient but not necessary. In Section 5–12, an exact criterion (Theorem 5–13) will be given for the consistency of any finite system of m linear equations in n variables.

EXERCISES

Solve the following systems of equations, using Cramer's Rule.

1. $\begin{aligned} x - 2y &= 3, \\ 2x - y &= 5. \end{aligned}$

2. $\begin{aligned} x + 2y - z &= 1, \\ x + y &= 5, \\ 3x - y + 2z &= 7. \end{aligned}$

3. $\begin{aligned} 3x - 4y + 2z &= 11, \\ x + 4y - 5z &= 12, \\ 5x + 2y + 3z &= 10. \end{aligned}$

4. $x + y + z + w = 5,$
$x - y \quad + 3w = 2,$
$y - 2z - w = 4,$
$x - y + z - w = 7.$

5. Prove Cramer's Rule for the system (5–25) as in Section 5–1 by multiplying each equation by the algebraic complement (in the matrix of the coefficients) of the coefficient of $x_1 = x$, adding the equations and solving for x_1. Repeat this process for the other variables.

6. Repeat Exercise 5 for (5–26).

7. Use the method of Exercise 5 to give a general proof of Theorem 5–12.

5–12 Systems of linear equations. In Section 5–11 we considered systems of n linear equations in n variables in which the determinant of the coefficients was different from zero. We now consider arbitrary finite systems of linear equations. Given a system of m linear equations in n variables

$$
\begin{aligned}
a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= b_1,\\
a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= b_2,\\
a_{31}x_1 + a_{32}x_2 + \cdots + a_{3n}x_n &= b_3,\\
&\cdot\\
&\cdot\\
&\cdot\\
a_{m1}x_1 + a_{m2}x_2 + \cdots + a_{mn}x_n &= b_m,
\end{aligned}
\tag{5–27}
$$

we define the *matrix of the coefficients* of the system to be

$$
\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n}\\
a_{21} & a_{22} & \cdots & a_{2n}\\
a_{31} & a_{32} & \cdots & a_{3n}\\
\cdot & \cdot & \cdots & \cdot\\
a_{m1} & a_{m2} & \cdots & a_{mn}
\end{bmatrix}
$$

and the *augmented matrix* of the system to be

$$
\begin{bmatrix}
a_{11} & a_{12} & \cdots & a_{1n} & b_1\\
a_{21} & a_{22} & \cdots & a_{2n} & b_2\\
a_{31} & a_{32} & \cdots & a_{3n} & b_3\\
\cdot & \cdot & \cdots & \cdot & \cdot\\
a_{m1} & a_{m2} & \cdots & a_{mn} & b_m
\end{bmatrix}.
$$

The system (5–25) has matrix and augmented matrix

$$
\begin{bmatrix} a_1 & b_1\\ a_2 & b_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_1 & b_1 & c_1\\ a_2 & b_2 & c_2 \end{bmatrix},
$$

respectively.

The rank of the augmented matrix of a system is always at least equal to the rank of the matrix of the coefficients, since every minor of the matrix of the coefficients is also a minor of the augmented matrix. For systems of n equations in n variables (Section 5–11), if the rank of the matrix of the coefficients is n, then the rank of the augmented matrix must also be n (since the augmented matrix has only n rows), and thus under the conditions stated in Theorem 5–12 the two matrices must be of the same rank. In general, we have

THEOREM 5–13. THE FUNDAMENTAL THEOREM FOR SYSTEMS OF LINEAR EQUATIONS. *A necessary and sufficient condition for a system of linear equations to be consistent is that the matrix of the coefficients be of the same rank as the augmented matrix.*

Theorem 5–13 and our next theorem, Theorem 5–14, are proved in [9] and other texts on determinants and matrices. We shall adopt them without proof and concern ourselves primarily with their applications. Since a set of linear equations in n variables is consistent if and only if the corresponding hyperplanes ($n > 3$), planes ($n = 3$), or lines ($n = 2$) represented by it have at least one point in common, we shall frequently make use of the geometric applications of these theorems (Exercises 6, 7, and 8).

THEOREM 5–14. *If in a system of linear equations in n variables the matrix of the coefficients and the augmented matrix are of the same rank r, then $n - r$ of the variables may be given arbitrary values and the remaining variables will then be uniquely determined.*

This theorem essentially states that the general solution of a system of linear equations in which both matrices have rank r contains $n - r$ parameters. It can also be shown (Exercise 12) that the general solution is linear in these parameters. The $n - r$ variables selected as parameters may be chosen in any manner such that the matrix of the coefficients of the remaining variables is of rank r.

Given the system of equations

$$
\begin{aligned}
x + y &= 2,\\
x - y &= 4,\\
2x + 2y &= 4,
\end{aligned}
$$

both the matrix of the coefficients and the augmented matrix are of rank two. Thus the system has a unique (since also $n = 2$) solution $x = 3$ and $y = 1$. The system

$$
\begin{aligned}
x + y &= 5,\\
x + y &= 2
\end{aligned}
$$

has augmented matrix of rank two, whereas the matrix of the coefficients is of rank one. Thus this system is inconsistent, i.e., there does not exist a pair of numbers satisfying both equations. The lines represented by this system of equations are parallel and distinct. Finally, both the matrices of the system

$$\begin{aligned} x + y &= 1, \\ 2x + 2y &= 2 \end{aligned}$$

are of rank one, the system is consistent, the two lines represented coincide, the value of one of the variables may be arbitrarily assigned (Theorem 5–14), and the general solution may be represented as $x = c$ and $y = 1 - c$ in terms of the parameter c.

Let us now consider the following example, in which $n = 3$. The system of equations

$$\begin{aligned} x \quad - z &= 1, \\ x + y \quad &= 2, \\ y + z &= 1 \end{aligned}$$

has coefficient matrix of rank two and augmented matrix of rank two. Since $n = 3$ and $r = 2$, one of the variables (in this case any one) may be used as a parameter. If z is taken as the parameter, the system becomes

$$\begin{aligned} x &= 1 + z, \\ x + y &= 2, \\ y &= 1 - z. \end{aligned}$$

Since the second equation is the sum of the other two, it may be discarded. The remaining two equations together with $z = z$ express the three variables of the given system in terms of the parameter z. There are unique values of the variables x and y corresponding to each value of z. The given system is satisfied by all points of the line $x - 1 = 1 - y = z$.

The above examples illustrate some of the applications of Theorems 5–13 and 5–14. Further applications will be considered in the following exercises and in the remaining three sections of this chapter.

Exercises

1. Prove that every system of linear homogeneous equations is consistent.

2. Prove that if a system of linear homogeneous equations has a unique solution, then that solution is $x_1 = x_2 = x_3 = \cdots = x_n = 0$.

3. Prove that any finite system of linear equations in n variables has a unique solution if and only if the ranks of the augmented matrix and the matrix of the coefficients are both equal to n.

4. Assume at least one $b_j \neq 0$ in (5–27) and prove that (a) if $m = n$, the nonvanishing of the determinant of the matrix of the coefficients is sufficient for a solution, (b) if $m = n + 1$, the vanishing of the determinant of the augmented matrix is a necessary condition for a solution.

5. Find the rank of the matrix of the coefficients, the rank of the augmented matrix, and all solutions (using parameters if necessary) of each of the following systems of equations:

(a) $\begin{aligned} 2x + 3y &= 6, \\ x - y &= 5. \end{aligned}$

(b) $\begin{aligned} x + 3y &= 3, \\ 2x + 6y &= 1. \end{aligned}$

(c) $\begin{aligned} 3x + 5y &= 2, \\ 6x + 10y &= 4. \end{aligned}$

(d) $\begin{aligned} x + y &= 3, \\ x - y &= 1, \\ 2x + y &= 5, \\ 2x - y &= 3. \end{aligned}$

(e) $\begin{aligned} x \quad - z &= 1, \\ x + y \quad &= 2, \\ y + z &= 1. \end{aligned}$

(f) $\begin{aligned} x + y + z &= 2, \\ x - y + z &= 1, \\ y \quad &= 1. \end{aligned}$

(g) $\begin{aligned} x + y + z &= 3, \\ x - y + z &= 1, \\ x \quad + z &= 4. \end{aligned}$

(h) $\begin{aligned} x + y + z &= 2, \\ x - y + z &= 1, \\ y \quad &= 5. \end{aligned}$

(i) $\begin{aligned} x \quad &= 1, \\ x + y \quad &= 2, \\ y + z &= 2. \end{aligned}$

6. Consider the systems of lines represented by the systems of equations in Exercise 5(a)–(d) and indicate which systems (a) intersect in a unique point, (b) represent coincident lines, (c) have no point in common.

7. Consider the systems of planes represented by the systems of equations in Exercise 5(e)–(i) and indicate which systems (a) intersect in a unique point, (b) intersect in a unique line, (c) coincide, (d) have no point in common.

8. Compare the results of Exercises 6 and 7 with those of Exercise 5 and discuss the geometric significance of Theorem 5–14.

9. Prove that if $m = n - 1$, $b_1 = b_2 = \cdots = b_m = 0$, and the matrix of the coefficients in (5–27) is of rank $n - 1$, then the ratios of the variables are

$$x_1 : x_2 : x_3 : \cdots : x_n = A_1 : -A_2 : A_3 : \cdots : (-1)^{n-1}A_n,$$

where A_j is the determinant obtained by deleting the jth column from the matrix of the coefficients [16; 41–42].

10. Apply the results of Exercise 9 to the following systems of equations:

(a) $\begin{aligned} x + y - z &= 0, \\ x - y + 2z &= 0. \end{aligned}$

(b) $\begin{aligned} 2x - 3y + z &= 0, \\ x - 3y + z &= 0. \end{aligned}$

11. Compare the results obtained in Exercise 10 with Theorem 5–14. Give a complete solution of each system in Exercise 10.

12. Prove that the general solution in Theorem 5–14 is linear in the $n - r$ parameters.

5–13 Linear dependence. In Section 5–9 we defined $c_1b_1 + c_2b_2 + \cdots + c_nb_n$, where the c's are constants not all zero and the b's are any elements to be a linear combination of the b's. This concept was used (Exercise 20, Section 5–9) in replacing each element, say a_{1j}, of one line of the matrix of a determinant by itself plus a linear combination of the corresponding elements on the remaining parallel lines, say,

$$a_{1j} + c_2a_{2j} + c_3a_{3j} + \cdots + c_na_{nj}, \quad (j = 1, 2, \ldots, n), \tag{5–28}$$

i.e., for all elements a_{1j} of the first row. For example, given the determinant

$$\begin{vmatrix} 1 & 2 & 3 \\ 3 & 2 & -2 \\ -1 & -6 & 1 \end{vmatrix},$$

we may replace the elements a_{1j} of the first row of its matrix by $a_{1j} + 2a_{2j} + a_{3j}$ $(j = 1, 2, 3)$ and obtain

$$\begin{vmatrix} 6 & 0 & 0 \\ 3 & 2 & -2 \\ -1 & -6 & 1 \end{vmatrix}$$

without changing the determinant (Theorem 5–9). Similarly, if we replace the elements a_{i1} of the first column of the matrix of

$$\begin{vmatrix} 1 & 11 & 6 & 9 \\ 4 & -3 & 0 & -1 \\ -2 & 7 & 3 & 5 \\ 3 & 6 & -3 & 3 \end{vmatrix}$$

by $a_{i1} + 3a_{i2} + 2a_{i3} - 5a_{i4}$, we obtain

$$\begin{vmatrix} 1 & 11 & 6 & 9 \\ 0 & -3 & 0 & -1 \\ 0 & 7 & 3 & 5 \\ 0 & 6 & -3 & 3 \end{vmatrix}.$$

In each of these examples we have used relations similar to (5–28) not only for a single set of numbers but for several sets of corresponding numbers. In the second example we used such a relation for

the four sets of corresponding numbers a_{1j}, a_{2j}, a_{3j}, a_{4j}, $(j = 1, 2, 3, 4)$, that is,

$$\begin{aligned} &a_{11} + 3a_{12} + 2a_{13} - 5a_{14}, \\ &a_{21} + 3a_{22} + 2a_{23} - 5a_{24}, \\ &a_{31} + 3a_{32} + 2a_{33} - 5a_{34}, \\ &a_{41} + 3a_{42} + 2a_{43} - 5a_{44}. \end{aligned}$$

Thus we considered the same linear combination for each of four sets of corresponding elements.

We now extend the concept of linear combination to that of linear dependence. The three sets of four numbers each:

$$\begin{matrix} x_1, & x_2, & x_3, & x_4, \\ y_1, & y_2, & y_3, & y_4, \\ z_1, & z_2, & z_3, & z_4, \end{matrix} \tag{5–29}$$

are said to be linearly dependent if there exist constants a, b, c not all zero such that

$$ax_j + by_j + cz_j = 0, \qquad (j = 1, 2, 3, 4).$$

Thus the three sets of numbers (5–29) are linearly dependent if and only if the system of linear homogeneous equations

$$\begin{aligned} ax_1 + by_1 + cz_1 &= 0, \\ ax_2 + by_2 + cz_2 &= 0, \\ ax_3 + by_3 + cz_3 &= 0, \\ ax_4 + by_4 + cz_4 &= 0, \end{aligned}$$

in which the x_j, y_j, z_j are given and the constants a, b, c are to be determined, has a solution with at least one of the constants different from zero. Accordingly, from Theorem 5–14 and Exercise 2, Section 5–12, the three sets of numbers (5–29) are linearly dependent if and only if the matrix of the coefficients

$$\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \\ x_4 & y_4 & z_4 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \end{bmatrix}$$

is of rank less than three, i.e., every third order determinant of the matrix is zero.

In general, m sets

$$a_{1j}, a_{2j}, \ldots, a_{nj} \qquad (j = 1, 2, \ldots, m) \tag{5–30}$$

of n elements each are said to be *linearly dependent* if and only if there exist constants $c_1, c_2, c_3, \ldots, c_m$ not all zero such that

$$
\begin{aligned}
&c_1a_{11} + c_2a_{12} + \cdots + c_ma_{1m} = 0,\\
&c_1a_{21} + c_2a_{22} + \cdots + c_ma_{2m} = 0,\\
&\cdot\\
&\cdot\\
&\cdot\\
&c_1a_{n1} + c_2a_{n2} + \cdots + c_ma_{nm} = 0,
\end{aligned}
\tag{5–31}
$$

that is, $c_1a_{i1} + c_2a_{i2} + \cdots + c_ma_{im} = 0 \quad (i = 1, 2, \ldots, n).$

The sets of elements (5–30) are *linearly independent* if the relations (5–31) imply $c_1 = c_2 = \cdots = c_m = 0$. The system of linear homogeneous equations (5–31) will be used in Theorem 5–15 to express necessary and sufficient conditions for the linear dependency of any m sets of elements (5–30) in terms of the matrix

$$
\begin{bmatrix}
a_{11} & a_{12} & \ldots & a_{1m}\\
a_{21} & a_{22} & \ldots & a_{2m}\\
\cdot & & & \\
\cdot & & & \\
\cdot & & & \\
a_{n1} & a_{n2} & \ldots & a_{nm}
\end{bmatrix}
\tag{5–32}
$$

of the coefficients of the c's. However, before considering this general case further, let us consider the special case (5–29) geometrically, assuming that the elements are real numbers.

Any three real numbers may be considered as coordinates of a point in Euclidean three-dimensional space. By definition, the three sets (5–29) are linearly dependent if and only if the four triples of corresponding numbers satisfy a relationship of the form

$$ax + by + cz = 0,$$

where a, b, c are constants not all zero, i.e. (under the assumption that the numbers are real), the triples of corresponding numbers are coplanar with the origin. There exist analogous and in a sense more elegant geometric interpretations of linear dependence in terms of homogeneous coordinates and in terms of vector spaces. We shall consider only the above more elementary interpretation in terms of nonhomogeneous coordinates in order to avoid the task of introducing other concepts.

Three sets of n real numbers each,

$$
\begin{aligned}
&x_1, \quad x_2, \quad \ldots, \quad x_n,\\
&y_1, \quad y_2, \quad \ldots, \quad y_n,\\
&z_1, \quad z_2, \quad \ldots, \quad z_n,
\end{aligned}
$$

are linearly dependent if and only if the n triples of corresponding numbers represent points coplanar with the origin. By Theorem 5–14, the system of equations

$$
\begin{aligned}
&ax_1 + by_1 + cz_1 = 0,\\
&ax_2 + by_2 + cz_2 = 0,\\
&\cdot\\
&\cdot\\
&\cdot\\
&ax_n + by_n + cz_n = 0
\end{aligned}
$$

has a unique solution $a = b = c = 0$ if the matrix

$$
\begin{bmatrix}
x_1 & y_1 & z_1\\
x_2 & y_2 & z_2\\
\cdot & & \\
\cdot & & \\
\cdot & & \\
x_n & y_n & z_n
\end{bmatrix}
\quad \text{or} \quad
\begin{bmatrix}
x_1 & x_2 & \ldots & x_n\\
y_1 & y_2 & \ldots & y_n\\
z_1 & z_2 & \ldots & z_n
\end{bmatrix}
\tag{5–33}
$$

is of rank three. The above system of equations has a solution in which at least one of the required constants a, b, c is different from zero if the matrix (5–33) is of rank less than three. Thus the above three sets of n numbers each are linearly dependent if and only if every third order determinant of the matrix (5–33) is zero. If the matrix is of rank two, the triples of corresponding numbers represent points coplanar with the origin; if the matrix is of rank one, they are collinear with the origin. The same conditions for linear dependence hold even though the geometric interpretation may not when the elements are from any ring.

The above discussion for three sets of n elements each may now be extended to any m sets of n elements each (5–30). When the elements are real numbers, each of the n sets of m corresponding real numbers may be taken as a point in Euclidean m-space. These n points are on a hyperplane

$$c_1a_1 + c_2a_2 + \cdots + c_ma_m = 0$$

through the origin if and only if the sets of elements (5–30) are linearly dependent. These conditions may be expressed as a system of equations (5–31) that apply whether the elements are real numbers or not. If $m = n$, the m sets of n elements each (5–30) are linearly dependent if and only if the matrix (5–32) is of rank less than m, that is, if and only if the m-rowed determinant of (5–32) is zero. If $m < n$,

then since the system (5–31) consists of n equations in m unknowns, the m sets of elements (5–30) are linearly dependent if and only if every m-rowed determinant of (5–32) is zero (Theorem 5–14). If $m > n$, there are not as many equations as there are constants to be determined and the sets are always dependent (Exercise 5). This situation is analogous to finding a plane $ax + by + cz = 0$ on one or two given points, as, for example, when the given sets are

$$\begin{matrix} x_1, & x_2, \\ y_1, & y_2, \\ z_1, & z_2. \end{matrix}$$

These results may be summarized as follows.

THEOREM 5–15. *If $m > n$, then any m sets of n elements each are linearly dependent. If $m \leqq n$, then m sets* (5–30) *of n elements each are linearly dependent if and only if every m-rowed determinant of the matrix* (5–32) *is zero.*

The definition that the elements of a set of constants $b_1, b_2, \ldots, b_n$ are linearly dependent if there exists a linear combination

$$c_1b_1 + c_2b_2 + \cdots + c_nb_n = 0$$

where not all the c's are zero may be extended to include arbitrary sets of elements. For example, whenever the variables take on values from an infinite set of numbers (Sections 3–1 and 3–4), m polynomials $f_1, f_2, \ldots, f_m$ in any number of variables are linearly dependent if and only if there exist m constants c_i not all zero such that

$$c_1f_1 + c_2f_2 + \cdots + c_mf_m = 0$$

for all values of the variables. This definition is equivalent (Exercise 10) to defining the polynomials of the set to be linearly dependent if and only if the sets of corresponding coefficients are linearly dependent. This alternate form of the definition may be used in Theorem 5–15. We shall consider several theorems based upon the definition of linear dependence and some of the many applications of this concept in the following set of exercises.

EXERCISES

1. Prove that if a set of elements is a linear combination of $m - 1$ other sets of elements, then the m sets of elements are linearly dependent.

2. Prove that if m sets of elements are linearly dependent, then at least one set is a linear combination of the others.

3. Prove that if there exist among m sets of elements k sets that are linearly dependent, where $k < m$, then the m sets are linearly dependent.

4. Prove that if any one of m sets of elements consists exclusively of zeros, then the m sets are linearly dependent.

5. Prove that for $m > n$ any m sets of n elements each are linearly dependent. (*Hint:* Extend the system to m sets of m elements each by adding zeros.)

6. Indicate which of the following sets, of four numbers each, are linearly dependent:

(a) $\begin{matrix} 3, & 0, & 1, & 5, \\ 1, & -2, & -1, & 2, \\ 2, & 2, & 2, & 3. \end{matrix}$

(b) $\begin{matrix} 2, & 2, & 1, & 3, \\ 3, & 5, & 2, & 4, \\ 1, & -1, & 0, & 2. \end{matrix}$

(c) $\begin{matrix} 1, & 0, & 1, & 1, \\ 0, & 1, & 1, & 0, \\ 1, & 1, & 0, & 0. \end{matrix}$

(d) $\begin{matrix} 1, & 2, & 3, & 4, \\ 2, & 4, & 6, & 8, \\ a, & b, & c, & d. \end{matrix}$

7. Prove that the three polynomials

$$a_jx + b_jy + c_jz + d_j \qquad (j = 1, 2, 3)$$

are linearly dependent if and only if the three sets of numbers

$$a_j, b_j, c_j, d_j \qquad (j = 1, 2, 3)$$

are linearly dependent.

8. Indicate which of the following sets of polynomials are linearly dependent:

(a) $3x + y + 2, \quad y - 1, \quad x + y + 2.$
(b) $x + 1, \quad y + 1, \quad x + y.$
(c) $x + 2y + 3z + 4, \quad 2x + 4y + 6z + 8, \quad ax + by + cz + d.$
(d) $x + 2y - z + 5, \quad 8z - 12y - 10, \quad 3x + z + 10.$

9. Prove that for any finite set of linearly dependent polynomials of the form $a_jx + b_jy + c_jz$ the graphs of the corresponding equations all intersect in at least one common point.

10. Prove that the linear dependence of any finite set of polynomials in any finite number of variables that take on values from an infinite set of numbers implies the linear dependence of the sets of constants comprising their sets of coefficients, and conversely.

5–14 Applications in analytic geometry. The use of determinants and matrices in geometry is recognized in a few elementary texts in analytic geometry and is a very important part (Section 5–15) of all advanced texts in geometry that use analytic methods. In this sec-

tion we shall extend the geometric concepts used in the study of linear dependence (Section 5–13) by simply enumerating some of the common applications of determinants and matrices in elementary analytic geometry. Then in the next section we shall conclude our study of determinants and matrices with a brief discussion of their applications to geometric transformations.

The area of a triangle with vertices at (x_1, y_1), (x_2, y_2), (x_3, y_3) is given by

$$\pm\,(\tfrac{1}{2})\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix},$$

where the sign is to be chosen so that the area is non-negative. This result may be extended to give

$$\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x_3 & y_3 & 1 \end{vmatrix} = 0$$

as a necessary and sufficient condition that the three points be collinear. It may also be used in the form

$$\begin{vmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ x & y & 1 \end{vmatrix} = 0$$

to find the equation of the line determined by two given distinct points (x_1, y_1) and (x_2, y_2).

In a plane two lines

$$a_1x + b_1y = c_1,$$
$$a_2x + b_2y = c_2$$

have a unique point in common if the matrices

$$\begin{bmatrix} a_1 & b_1 \\ a_2 & b_2 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \end{bmatrix}$$

are both of rank two, coincide if the matrices are both of rank one, and are parallel if the matrices are of different ranks (Theorems 5–13 and 5–14). Similarly, in three-dimensional space two planes

$$a_1x + b_1y + c_1z = d_1,$$
$$a_2x + b_2y + c_2z = d_2$$

have a unique line in common if in this system of equations the matrix of the coefficients and the augmented matrix are both of rank two,

coincide if the matrices are both of rank one, and are parallel if the matrices are of different ranks. The concepts in Section 5–12 may also be used to prove that three planes

$$a_jx + b_jy + c_jz = d_j \qquad (j = 1, 2, 3)$$

have a unique point in common if the corresponding matrices are both of rank three, have a unique line in common if the matrices are both of rank two, coincide if the matrices are both of rank one, and have no point in common if the matrices are of different ranks. Further correspondences between the planes and the ranks of the matrices are considered in Exercise 7.

The volume of a tetrahedron with vertices (x_1, y_1, z_1), (x_2, y_2, z_2), (x_3, y_3, z_3), (x_4, y_4, z_4) is given by

$$\pm\,(\tfrac{1}{6})\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{vmatrix},$$

where, as before, the sign is to be chosen so that the volume is non-negative. Also as before, this result may be extended to give

$$\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x_4 & y_4 & z_4 & 1 \end{vmatrix} = 0$$

as a necessary and sufficient condition that the four given points be coplanar. Three points in space (x_j, y_j, z_j), $j = 1, 2, 3$, are noncollinear if and only if

$$\begin{vmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{vmatrix} \neq 0,$$

and the equation of the plane determined by three given noncollinear points is given by

$$\begin{vmatrix} x_1 & y_1 & z_1 & 1 \\ x_2 & y_2 & z_2 & 1 \\ x_3 & y_3 & z_3 & 1 \\ x & y & z & 1 \end{vmatrix} = 0.$$

We have now seen that the equation of the line determined by any two distinct points and the equation of the plane determined by three noncollinear points may be expressed in terms of determinants. Also

the area of a triangle and the volume of a tetrahedron may be expressed in terms of the coordinates of their vertices and determinants. Such examples illustrate the application of determinants and matrices in analytic geometry. A few other examples will be considered in the following exercises; many examples may be found in [16].

EXERCISES

1. Indicate which of the following triples of points in a plane are collinear. If they are not collinear, find the area of the triangle determined by them:
   (a) (1, 2), (5, 6), (17, 18).
   (b) (−1, 5), (1, 4), (3, 0).
   (c) (11, 7), (6, 2), (−1, 3).

2. Use determinants to indicate the equations of the lines determined by the following pairs of points:
   (a) (1, 2), (5, 6).
   (b) (−1, 5), (1, 4).
   (c) (11, 7), (6, 2).
   (d) (912, −13), (−115, 76).

3. Indicate which of the following sets of points in space are coplanar. If they are not coplanar, find the volume of the tetrahedron determined by them:
   (a) (1, 2, 3), (4, 5, 6), (7, 8, 9), (10, 11, 12).
   (b) (2, −2, 0), (5, 7, 11), (−7, 3, 12), (1, 1, 1).
   (c) (1, −1, 1), (7, 13, 27), (5, 2, 1), (−6, 3, 4).

4. Use determinants to indicate the equation of the plane on the four points or the equations of the faces (planes) of the tetrahedron in (a) Exercise 3(a), (b) Exercise 3(b), (c) Exercise 3(c).

5. Describe the graphs in the plane of the sets of equations in Exercise 5(a)–(d), Section 5–12.

6. Describe the graphs in space of the sets of equations in Exercise 5(e)–(i), Section 5–12.

7. Give the ranks of the matrix of the coefficients and the augmented matrix of a system of equations representing
   (a) three planes having a unique point in common,
   (b) three distinct planes having a line in common,
   (c) three planes having a plane in common,
   (d) three parallel planes,
   (e) two parallel planes and a third plane intersecting them,
   (f) two coincident planes and a third plane intersecting them,
   (g) two coincident planes and a third plane parallel to them,
   (h) three distinct planes such that the pairs of planes intersect in three parallel lines.

8. Prove [16; 87] that the line determined by the intersecting planes

$$a_1x + b_1y + c_1z + d_1 = 0,$$
$$a_2x + b_2y + c_2z + d_2 = 0$$

has direction numbers

$$\begin{vmatrix} b_1 & c_1 \\ b_2 & c_2 \end{vmatrix}, \quad \begin{vmatrix} c_1 & a_1 \\ c_2 & a_2 \end{vmatrix}, \quad \begin{vmatrix} a_1 & b_1 \\ a_2 & b_2 \end{vmatrix}.$$

(*Hint:* Use Exercise 9, Section 5–12. Direction numbers are defined in analytic geometry texts that include solid geometry.)

9. Apply the result of Exercise 8 to the lines determined by
   (a) $x + y - z + 2 = 0$, $x - y + 2z - 5 = 0$,
   (b) $2x + 3y - z - 3 = 0$, $x - 5y + z + 2 = 0$.

5–15 Geometric transformations. The importance of transformations in geometry is indicated by Felix Klein's definition: Any geometry is a study of properties (expressed by definitions and theorems) left invariant under a group of transformations. For example, in Euclidean geometry we study properties such as length, area, magnitude of angles, parallel lines, similar and congruent triangles that remain invariant under rigid motions, i.e., translations and rotations. Each of these transformations may be expressed as a matrix with reference to a coordinate system.

Given the ordinary xy-plane used in analytic geometry, we may represent any translation in the plane by the system of equations

$$x' = x + a, \quad y' = y + b. \tag{5–34}$$

For example, using the axes in the conventional positions, if every point is moved two units to the right, we have

$$x' = x + 2, \quad y' = y;$$

if every point is moved three units down, we have

$$x' = x, \quad y' = y - 3;$$

if every point is moved two units to the right and three units down, we have

$$x' = x + 2, \quad y' = y - 3.$$

In general, since any translation in the plane may be considered as the result of a motion along the x-axis and a motion along the y-axis, we have (5–34) for any translation in the plane. Similarly, one proves in analytic geometry that any rotation about the origin in the plane may be expressed in the form

$$x' = x \cos\theta - y \sin\theta, \quad y' = x \sin\theta + y \cos\theta. \tag{5–35}$$

Also one may prove that if the rotation (5–35) is followed by the translation (5–34), we have a transformation of the form

$$x' = x\cos\theta - y\sin\theta + a, \quad y' = x\sin\theta + y\cos\theta + b.$$

We now consider methods of indicating these transformations by means of matrices.

Each of the above transformations is represented by equations of the form

$$x' = a_{11}x + a_{12}y + a_{13}, \quad y' = a_{21}x + a_{22}y + a_{23}.$$

Furthermore, each transformation is completely determined by the a_{jk}'s, that is, by a matrix of the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{bmatrix}.$$

We extend this matrix and use a third order matrix of the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{bmatrix},$$

obtained by adding the third row of the identity matrix (Exercise 6) so that the matrices of two transformations may be multiplied (Section 5–10) and the matrix of the product will have the same form as the given matrices. We shall find that an ordered product of the matrices of two transformations is the matrix of a transformation obtained by applying the two given transformations one after the other in a certain order. This property underlies many of the exercises at the end of this section.

We have now seen that any translation (5–34) and any rotation about the origin (5–35) may be respectively represented by the matrices

$$\begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 1 \end{bmatrix} \text{ and } \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Similarly, any point (x, y) in the plane may be represented by a matrix. Also as before, the form of the matrix is chosen to allow multiplication of certain matrices. We shall use a matrix with one column and three rows

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix}.$$

This convention will enable us to express the equations of a transformation as a single equality in terms of matrices.

Two matrices are said to be *equal* if and only if they have the same number of rows, the same number of columns, and corresponding elements are identical. Two matrices related by elementary transformations (Exercise 16, Section 5–10) are of the same rank but are not necessarily equal in the sense considered here. The equality of two matrices of mn elements is equivalent under the above definition to a system of mn equations. Thus by multiplying matrices, we may express the translation (5–34) by the equation

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & a \\ 0 & 1 & b \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} x+a \\ y+b \\ 1 \end{bmatrix}.$$

Since the first and last matrices are equal if and only if corresponding elements are equal, the above equality of matrices is precisely equivalent to (5–34). Similarly, (5–35) is equivalent to

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}.$$

The relationship between the expression of a translation or rotation in terms of a system of linear equations in the coordinates and in terms of a matrix (in one sense the coefficient matrix of the system of equations) may be quickly grasped (Exercises 1 and 2), so that one representation is as easy to obtain as the other. One advantage of the matrix representation arises from the ease with which the result of a sequence of transformations may be obtained as the product (taken in reverse order) of the corresponding matrices. The translation obtained as a sequence of two translations considered above in the explanation of a translation may be expressed as

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 \\ 0 & 1 & -3 \\ 0 & 0 & 1 \end{bmatrix}.$$

In the special case of two translations, the order of multiplying the matrices is unimportant (Exercise 5), but in general we shall find that order must be considered.

The transformation obtained as the result of considering two transformations in order is called the *ordered product* of those two trans-

formations. Similarly, we may consider the ordered product of any finite number of transformations. The importance of order is indicated by the fact that if the translation (5–34) is followed by the rotation (5–35), one obtains

$$\begin{bmatrix} \cos\theta & -\sin\theta & a\cos\theta - b\sin\theta \\ \sin\theta & \cos\theta & a\sin\theta + b\cos\theta \\ 0 & 0 & 1 \end{bmatrix}, \tag{5–36}$$

whereas if the rotation is taken first and followed by the translation, we have

$$\begin{bmatrix} \cos\theta & -\sin\theta & a \\ \sin\theta & \cos\theta & b \\ 0 & 0 & 1 \end{bmatrix}. \tag{5–37}$$

These results may be readily verified by multiplying the matrices of the transformations in the opposite order from that in which the transformations are used (Exercise 3).

The fact that (5–36) and (5–37) are in general different indicates that the effect of a translation followed by a rotation is different from that of the rotation followed by the translation. For example, consider the translation $x' = x + 2$, $y' = y$, and the rotation $x' = -x$, $y' = -y$. The point (3, 6) is taken to the point (5, 6) under the translation (assuming that the coordinate system remains fixed) and the point (5, 6) is taken to (−5, −6) under the rotation, i.e., the translation followed by the rotation takes (3, 6) to (−5, −6). Similarly the point (3, 6) is taken to the point (−3, −6) by the rotation and the point (−3, −6) is taken to (−1, −6) by the translation, i.e., the rotation followed by the translation takes (3, 6) to (−1, −6). Thus we find that the application of transformations and the multiplication of matrices are noncommutative operations (Exercise 4).

The transformations (3–34) to (3–37) of Euclidean geometry may be considered as special cases of the transformations of more general geometries. Any transformation (affine transformation) of the Euclidean plane into itself [52; 117–118] may be represented by a matrix of the form

$$\begin{bmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5–38}$$

where $a_1b_2 - a_2b_1 \neq 0$. If $b_1 = a_2 = 0$ and $a_1 = b_2 = 1$, then (5–38) represents a translation; if $c_1 = c_2 = 0$, $a_1 = b_2$, $a_2 = -b_1$ and $a_1^2 + a_2^2 = 1$, then (5–38) represents a rotation. Any rigid motion in the plane (Euclidean transformation) may be expressed as the product of a translation and a rotation (possibly in space) and may be represented by a matrix of the form

$$\begin{bmatrix} a & b & c_1 \\ -be & ae & c_2 \\ 0 & 0 & 1 \end{bmatrix}, \tag{5–39}$$

where $a^2 + b^2 = 1$ and $e^2 = 1$ (Exercises 13, 14). When $e = 1$, (5–39) is equivalent to (5–37) (Exercise 15); when $e = -1$, a rotation in space or a line reflection in the plane must be included.

Since any geometry may be considered as a study of properties invariant under groups (Exercise 9) of transformations and these transformations may be represented by matrices, this section could be expanded into a whole book. We have simply indicated how two common transformations could be expressed in terms of matrices (5–34) and (5–35) and indicated that these transformations are merely special cases of more general transformations such as (5–38) in geometries that include the Euclidean geometry as a special case. A thorough study of the manner in which Euclidean geometry is a special case of several other geometries may be found in [35].

In this chapter we have defined determinants of square matrices of any order n in terms of permutations and have discussed the use of determinants and matrices in the study of systems of linear equations, linear dependence, analytic geometry, and geometric transformations. Our discussion has, of necessity, been restricted to a small number of typical applications. More complete discussions of the theory and applications may be found in texts such as [9], [16], [39], [44], and [49].

Exercises

1. Use matrices to represent the following translations:
   (a) $x' = x - 1$, $y' = y + 2$,
   (b) $x' = x + 2$, $y' = y + 5$,
   (c) $x' = x - 3$, $y' = y - 4$.
2. Use matrices to represent rotations about the origin of (a) 30°, (b) 45°, (c) 120°, (d) 180°, (e) 270°.
3. Derive the matrices (5–36) and (5–37) from (5–34) and (5–35).
4. Use the matrices

$$\begin{bmatrix} a & b & c \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} d & e & f \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

to illustrate the fact that multiplication of matrices is noncommutative.

5. Prove that the product of any two translations is (a) a translation, (b) commutative.

6. Prove that

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

represents the identity transformation on the plane, i.e., when it multiplies or is multiplied by any other matrix of three rows and three columns, the product is the same as the other matrix.

7. Write the products of the transformations found in the following exercises as single transformations, using matrices: (a) Exercises 1(a) and 2(a); (b) Exercises 1(a) and 1(b); (c) Exercises 1(b) and 2(c); (d) Exercises 1(c) and 2(d); (e) Exercises 1(c) and 2(e); (f) Exercises 2(a) and 2(d).

8. Use the result of Exercise 6 and show that the translation $x' = x - a$, $y' = y - b$ is the inverse of (5–34).

9. A set of affine transformations (5–38) forms a group if the set contains the inverse of every transformation of the set and the product of every pair of transformations of the set. Prove that this definition is in accord with the general definition of a group given in Section 1–14.

10. Prove that the set of all translations (5–34) forms a group.

11. Prove that the set of all rotations about the origin forms a group.

12. Prove that the set of all affine transformations (5–38) forms a group.

13. Prove that (5–34) and (5–35) each have the form (5–39), where $a^2 + b^2 = e^2 = 1$.

14. Prove that (5–36) and (5–37) each have the form (5–39), where $a^2 + b^2 = e^2 = 1$.

15. Prove that (5–39) may be put in the form (5–37) when $a^2 + b^2 = e = 1$.

16. Prove that a translation is determined by one pair of corresponding points.

17. Any *point reflection* may be expressed in the form

$$x' = -x + a, \quad y' = -y + b.$$

Give the corresponding representation of a point reflection by a matrix. Does the set of all point reflections form a group? Explain.

18. Prove that the product of an even number of point reflections is a translation.

19. Prove that the product of an odd number of point reflections is a point reflection.

20. Prove that the set of all point reflections and translations forms a group.

21. Any *dilation* may be expressed in the form

$$x' = ax + b, \quad y' = ay + c, \qquad \text{where } a \neq 0, 1.$$

Give the corresponding representation of a dilation by a matrix. Does the set of all dilations form a group? Explain.

22. The set of all dilations and translations constitutes the set of *homothetic transformations*. Prove that the set of all homothetic transformations forms a group.

23. Prove that the set of all matrices of the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

with nonvanishing determinants forms a group.

24. Prove that the set of all matrices of the form

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}$$

with nonvanishing determinants forms a group.

CHAPTER 6

CONSTRUCTIONS

Geometric constructions have an appeal to young and old. Children enjoy making table decorations for their birthday parties, Christmas, or other special events. Probably the pleasure of making paper baskets, pasting on the handles, and coloring the sides is enhanced by dreams of having the basket full of candy at the party. As the child grows older the appeal of geometric constructons may include playing with rulers, compasses, and protractors; making stars, making valentines, and later making geometric solids. Some adults turn from paper and paste to wood carving or metal work. Others make model trains, boats, and even complete model towns. A few decide to challenge mathematical authorities and try to solve the classical problem of trisecting an angle (Sections 6–6 to 6–8). To young and old the fascination of geometric constructions may bring many happy hours.

Throughout this chapter we shall be concerned with constructions in Euclidean plane geometry. We shall find (Section 6–4) that classical constructions using only straightedge and compasses may be used to perform the four rational operations and the extraction of square roots. Thus, given any line with an origin and a unit point to designate the unit of distance and positive sense or direction along the line, we may associate positive integers with points obtained by adding units along the line, negative integers with points obtained by subtracting units, rational numbers with points obtained by multiplication and division. In other words, we may construct the set of rational points with respect to the given origin and unit point on the line. In general, we may construct with respect to the given points all points on the line with coordinates expressible in terms of a finite number of rational numbers, rational operations, and extraction of square roots. Furthermore, these are the only points on the line that may be constructed from the given points using only straightedge and compasses (Section 6–3). Thus there exist precise algebraic criteria for determining whether or not a given point on a line may be constructed from the given origin and unit of distance, using classical methods. Similarly, there exist algebraic criteria for de-

termining the constructibility of plane figures. These algebraic criteria provide the basis for our present consideration of geometric constructions in this study of fundamental concepts of algebra. We shall analyze classical constructions in the plane from an algebraic viewpoint (Sections 6–3 and 6–4), use our algebraic concepts to prove the impossibility of three famous classical construction problems (Section 6–6), and consider several nonclassical constructions (Sections 6–7 and 6–8) of one of these problems, the trisection of arbitrary angles.

This study of geometric constructions from an algebraic viewpoint provides a slight insight into the underlying relations between algebra (considered as a study of sets of numbers and variables and relations among them) and geometry (considered as a study of sets of points, lines, planes, etc. and relations among them). At the foundations of mathematics there are basic theories and concepts that apply equally well to algebra and to geometry. We have not endeavored to reach this common core of all mathematics. However, it is hoped that the consideration of the rational operations both algebraically and geometrically in this chapter and the geometric representation of certain common algebraic functions in the next chapter will give the reader some appreciation of the interdependence of algebra and geometry.

6–1 Classical constructions. Geometric constructions may be classified in two sets, according to the methods and apparatus used. The early Greeks endeavored to make all elementary geometric constructions using only compasses and a straightedge. The straightedge may be used only to draw straight lines. One is not allowed to use the length of the straightedge, the width of the straightedge, or marks upon the straightedge. We shall refer to constructions made under these restrictions as *classical constructions*. Constructions made using marked rulers, protractors, linkages, etc. (Section 6–9) will be called *nonclassical constructions*.

Theoretically, constructions by straightedge and compasses are absolutely accurate. However, for all practical purposes they are neither more nor less accurate than constructions with protractor and marked ruler. Many of the problems considered difficult or impossible under the classical restrictions are simple when we use marks on the straightedge, protractor, parallel rulers, pantograph, angle trisector, linkages, and similar devices. Birkhoff and Beatley consider constructions with ruler and protractor [6; 165–171] as well as con-

structions with straightedge and compasses [6; 172–196]. Fourrey [20] considers several types of constructions, including constructions with straightedge only, constructions with compasses only, and constructions with straightedge and compasses.

For both the classical and the nonclassical constructions we assume that all points, lines, etc. are in the same Euclidean plane. For the classical constructions we make the following five fundamental assumptions:

(i) A straight line may be drawn through any two given points.

(ii) A circle may be drawn with any given point as center and any given line segment as radius.

(iii) The intersection of any two given nonparallel lines may be determined.

(iv) The intersections of any given line and any given circle may be determined if such exist.

(v) The intersections of any two given circles may be determined if such exist.

Every classical construction must consist of a finite number of steps with straightedge and compasses. Since the above five assumptions include all possible steps with straightedge and compasses, every classical construction must consist of a finite number of steps, where each step depends upon one of the above assumptions.

Exercises

1. State algebraic assumptions equivalent to each of the above five fundamental assumptions for classical constructions.

2. Make the following constructions, using only a straightedge [20; 3–24]:

(a) Given a line segment AB and a line m parallel to AB, find the mid-point of the segment AB.

(b) Given two parallel lines and a point P, construct a line through P parallel to the given lines.

(c) Given the lines $x = 0, x = 1, y = 0, y = 1$ of a coordinate system on a plane, make or describe constructions for the points $(2, 0)$, $(3, 0)$, $(k, 0)$, $(0, 2)$, $(0, 3)$, $(0, n)$, (k, n), where k and n are any integers.

3. Make the following constructions, using only compasses [20; 95–114]:

(a) Given three points in a plane, determine whether or not they are collinear.

(b) Given a line m with a segment AB marked upon it, construct a segment AF on m of length five times that of AB.

(c) Given any two points A and B, construct, without considering the line AB, a point C collinear with AB and such that the length of AC is twice that of AB.

(d) Given the points $(0, 0)$, $(0, 1)$, $(1, 0)$, illustrate and describe a construction for any integral point (k, n) as in Exercise 2(c).

4. Mascheroni discovered in the eighteenth century that any point that could be constructed using straightedge and compasses could also be constructed using only compasses. Prove that all but the first of the above five fundamental constructions assumed for classical constructions may be performed using only compasses [13; 140–152].

6–2 Elementary classical constructions. Ever since the time of the early Greeks, geometers have been fascinated by the large number of constructions that may be made using only straightedge and compasses, i.e., classical constructions. At present most high school geometry texts include some elementary classical constructions. In particular, [6; 172–196] contains an excellent presentation of classical constructions, including those listed as exercises at the end of this section. These exercises will serve primarily as a review of the common constructions considered in secondary school. Hints are given in a few cases. Some of these constructions may be greatly simplified by using advanced theorems of Euclidean geometry. It is also a good supplementary exercise to give geometric or algebraic proofs of each construction. Although the list of exercises is long and could have been much longer, there are also many constructions that are not possible using only straightedge and compasses. In the next three sections we shall consider an algebraic basis for stating whether or not a specified construction is possible using only straightedge and compasses.

Exercises

Construct the following, using only straightedge and compasses:

1. The perpendicular bisector of a given line segment.
2. An angle equal to a given angle.
3. The bisector of a given angle.
4. A line through a given point and parallel to a given line.
5. The division of a given line segment into n equal parts.
6. The division of a given line segment into parts proportional to k given line segments.
7. The fourth proportional to three given line segments, that is, n for $r/s = m/n$.
8. The perpendicular to a given line at a given point (a) on the line, (b) not on the line.

9. The mean proportional (geometric mean) between two given line segments, that is, s for $m/s = s/n$.

10. The circle through three given noncollinear points.

11. The circle circumscribed about a given triangle.

12. The circle inscribed in a given triangle.

13. The center of a circle given an arc.

14. The tangent at a given point on a given circle.

15. The tangents to a given circle from a given external point.

16. The common external tangents of two given circles when such exist. (*Hint:* Given circles with centers O, O' of radii r, r' respectively, where $r \geqq r'$, construct a circle about O of radius $r - r'$ and use Exercise 15, where the external point is taken as O'.)

17. The common internal tangents of two given circles when such exist. (*Hint:* This construction may be done in a manner similar to that used in Exercise 16 by first constructing a circle of radius $r + r'$ about O.)

18. A triangle having three sides given.

19. A triangle having two sides and the included angle given.

20. A triangle having two angles and a side given. (*Hint:* Given any two angles of a triangle, the third angle may be found using the fact that in Euclidean geometry the sum of the three angles of a triangle is a straight angle.)

21. A triangle (not always unique) having two sides and an angle opposite one of them given.

22. The trisection of a right angle.

23. Regular polygons of three, six, and twelve sides inscribed in a given circle.

24. Regular polygons of four, eight, and sixteen sides inscribed in a given circle.

25. Regular polygons of five and ten sides inscribed in a given circle.

26. A regular polygon of fifteen sides inscribed in a given circle. [*Hint:* Use a side or its central angle from a regular inscribed hexagon (six sides) and the same from a regular inscribed decagon (ten sides).]

27. Regular polygons of three, four, five, six, eight, ten, and twelve sides circumscribed about a given circle.

6–3 The algebraic viewpoint. The Greek geometers devised many constructions using straightedge and compasses. However, they tried in vain to solve by classical constructions such problems as the duplication of a cube, the quadrature of a circle, and the trisection of an angle (Section 6–6). During the nineteenth century, algebraic proofs were given to show that these three problems cannot be solved using only straightedge and compasses (Section 6–6). They can be solved, however, by using nonclassical constructions. In general, the algebraic or analytic criterion for constructibility (Sections 6–4 and 6–5) enables one to determine exactly which classical construction problems are solvable, i.e., which construction problems are solvable using only straightedge and compasses. For example, algebraic considerations led to the classical construction of a regular polygon of seventeen sides whose constructibility under the classical restrictions was not even suspected during the twenty centuries from the time of Euclid to that of Gauss [15; 353].

The algebraic statements of the five assumptions stated in Section 6–1 are:

(i′) The equation of a straight line through any two given points may be determined.

(ii′) The equation of a circle with given center and radius may be determined.

(iii′) The coordinates of the point of intersection of any two given nonparallel lines may be determined.

(iv′) The coordinates of the points of intersection of a given line and a given circle may be determined if such exist.

(v′) The coordinates of the points of intersection of two given circles may be determined if such exist.

Each of the above results may be found algebraically using rectangular Cartesian coordinates, the four rational operations, and the extraction of real square roots. Conversely, using straightedge and compasses, only problems that are algebraically equivalent to the above can be solved. Therefore algebraic criteria exist for determining whether or not any specified construction may be accomplished using only straightedge and compasses. These criteria are most easily stated in terms of the four rational operations and the extraction of square roots. In the following section we shall consider specific classical constructions, the *basic classical constructions*, that may be used to perform these five operations.

6–4 Basic classical constructions. We have just pointed out that every possible classical construction must be algebraically equivalent to a finite number of steps involving only the four rational operations and the extraction of square roots. We now verify that these five operations may be accomplished using only straightedge and compasses.

Given line segments of length m and n respectively, we may easily

construct segments of length m, n, $m+n$, and $m-n$ on any given line. If we are also given a segment of unit length, we may construct segments of length mn and m/n as in Fig. 6–1. These two constructions make use of the fact that a line parallel to one side of a triangle

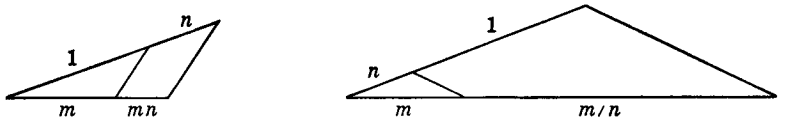


FIG. 6–1

divides the other two sides in the same ratio. Thus we have the proportions $mn : m = n : 1$ and $m : (m/n) = n : 1$ respectively from the triangles in Fig. 6–1. We may now add, subtract, multiply, and divide line segments in the above sense using only straightedge and compasses, i.e., the four rational operations may be performed as classical constructions.

The extraction of a square root is a special case of finding the geometric mean or mean proportional (Exercise 9, Section 6–2). The proof of this classical construction is based upon the fact that any triangle inscribed in a semicircle is a right triangle. Thus $\triangle ABC$ in Fig. 6–2 is a right triangle. The segment CD is perpendicular to AB at D, where $AD = m$ and $DB = n$. Then triangles ADC and CDB are similar and $AD/CD = CD/DB$, whence $CD = \sqrt{mn}$. The special case $\sqrt{m}$ in which we are primarily interested may be accomplished (take $n = 1$) for any given line segment of length m by classical methods whenever a segment of unit length is given (or may be obtained by classical methods from the given data).

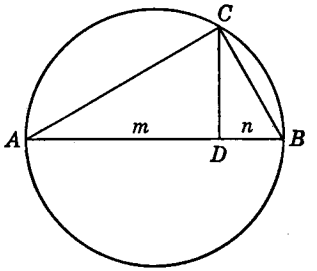


FIG. 6–2

Given a segment of unit length and segments m and n, we may construct segments $m+n$, $m-n$, $m \cdot n$, m/n and $\sqrt{mn}$, that is, given a segment of unit length we have now verified that any given line segments may be combined under the four rational operations and the extraction of square roots using only straightedge and compasses. Conversely, since every classical construction must consist of a finite number of applications of the five fundamental assumptions (Section 6–1) and each of these may be performed using the four rational operations and the extraction of square roots, we have proved that every classical construction must consist of a finite set of the basic classical constructions. We now define a *geometric plane figure* as any set of points and lines in a plane, and we have

THEOREM 6–1. *A geometric plane figure is constructible by straightedge and compasses if, and only if, the rectangular Cartesian coordinates of its points (vertices, etc.) can be derived from those of the given figure by a finite number of rational operations and extractions of real square roots.*

All rational numbers and expressions such as $\sqrt{10 - 2\sqrt{5}}$ can be constructed by straightedge and compasses as soon as a unit is chosen. The above expression is the length of a side of a regular pentagon inscribed in a circle of radius two. In the next section we shall see that every such constructible expression is a root of an irreducible equation with integral coefficients having as its degree an integral power of 2.

EXERCISES

1. Given a unit segment, construct segments of lengths 3, $\frac{3}{4}$, 2, $2 + \sqrt{5}$, $\sqrt{3 + \sqrt{2}}$.

2. Divide a given line segment into five equal parts.

3. Divide a given line segment into parts proportional to three given line segments.

4. Construct a five-pointed star.

5. Construct a regular hexagon when its side is given.

6. Given an obtuse triangle, construct the inscribed and circumscribed circles.

7. Give an algebraic proof of the construction of a regular decagon [13; 122–123], [6; 191–194].

8. Given the x-axis, origin and point (1, 1) in a plane, construct points: (a) (4, 0), (b) $(-\frac{3}{2}, 0)$, (c) $(-5 + \sqrt{2}, 0)$, (d) (2, 3), (e) $(\sqrt{3}, \sqrt{5})$.

9. As in Exercise 8, construct the graphs of (a) $x = 5$, (b) $2x + 3y = 6$, (c) $\sqrt{3}x + 4y = 2\sqrt{3}$.

10. Prove that the graph of any line with constructible coefficients may be constructed by classical methods.

11. Prove that any circle

$$x^2 + y^2 + bx + dy + e = 0$$

with constructible coefficients may be constructed.

6–5 Construction of roots of equations. Any linear equation has a root that may be expressed in terms of the coefficients, using only rational operations. Thus any linear equation with constructible coefficients has a constructible root where a number is said to be *constructible* if a corresponding line segment may be obtained from the given data under whatever restrictions are being considered. Throughout this section we shall be concerned with numbers that are constructible using only straightedge and compasses.

Any quadratic equation may be put in the form

$$x^2 - ax + b = 0. \tag{6–1}$$

The roots of this equation are precisely [15; 355–356] the intersections with the x-axis of the circle

$$\left(x - \frac{a}{2}\right)^2 + \left(y - \frac{b+1}{2}\right)^2 = \frac{a^2 + (b-1)^2}{4}, \tag{6–2}$$

which reduces to

$$\left(x - \frac{a}{2}\right)^2 = \frac{a^2}{4} - b$$

when $y = 0$. Any quadratic equation with constructible coefficients may be put in the form (6–1), where a and b are constructible. The coordinates of the center and the radius of the circle (6–2) are then constructible. Thus the real roots of any quadratic equation with constructible coefficients may be constructed whenever such real roots exist. We have now proved

THEOREM 6–2. *If the coefficients of a linear or a quadratic equation may be constructed from the given data using only straightedge and compasses, then the real roots of the equation may be constructed using only straightedge and compasses.*

We next consider a few results for equations of degree greater than two. However, we shall not attempt to give a complete theory for these equations.

In the ring of polynomials with integral coefficients (Section 3–2) a given polynomial $p(x)$ is reducible over the ring of integers (Section 3–6) if and only if it may be expressed in the form $p(x) = q(x) \cdot r(x)$, where $q(x)$ and $r(x)$ are polynomials of positive degree with integral coefficients. In particular, any nonlinear polynomial with integral coefficients and a rational root is reducible over the ring of integers. If an irreducible equation $f(x) = 0$ with rational coefficients has a

constructible root r, it has a root that may be expressed using the rational operations and the extraction of square roots. Then (from more advanced theories of algebra) all the roots of $f(x) = 0$ are conjugates of r and may be obtained from r just as $1 - \sqrt{2}$ may be obtained from $1 + \sqrt{2}$. For example, the number $\sqrt{10 - 2\sqrt{5}}$ mentioned at the end of Section 6–4 is a root of $x^4 - 20x^2 + 80 = 0$. This equation has roots $\sqrt{10 - 2\sqrt{5}}$, $\sqrt{10 + 2\sqrt{5}}$, $-\sqrt{10 - 2\sqrt{5}}$, $-\sqrt{10 + 2\sqrt{5}}$. In general, the degree d of any irreducible equation having a constructible real root must be of the form $d = 2^k$, where k is a non-negative integer. This is intuitively evident, since any rational root is a root of an irreducible equation of degree $1 = 2^0$ and any constructible surd root is one of an even number (including itself) of conjugate roots obtained from the given root by considering each of the radicals in turn as positive and negative. Let $f(x) = 0$ be a polynomial equation with rational coefficients having precisely the 2^m conjugates obtained in this manner as roots. Then $f(x)$ has degree 2^m. This method of counting conjugate roots may, however, cause some roots to be counted more than once as, for example, when the given constructible number is

$$\sqrt{5 - \sqrt{2} + \sqrt{3}} + \sqrt{5 + \sqrt{2} - \sqrt{3}}.$$

In such cases it can be shown [30; 5–12] that every root is counted exactly s times for some positive integer s where s divides m, and that if $g(x) = 0$ is an irreducible polynomial equation with rational coefficients having the given constructible number as a root, then $g(x) = 0$ has as its roots the distinct roots of $f(x) = 0$ and $f(x) = c[g(x)]^s$, where c is a constant. Thus the degree d of $g(x)$ satisfies the relation $d^s = 2^m$, whence $d = 2^k$ for some positive integer k. As a consequence of this result, we have

THEOREM 6–3. *An irreducible polynomial equation with rational coefficients and of degree d, where d cannot be expressed as an integral power of 2, has no root that may be constructed from the unit distance using only straightedge and compasses.*

Note that one cannot construct all real roots of irreducible polynomial equations with rational coefficients and degree 2^m for every integer m (Section 4–5). We shall find Theorem 6–3 very useful when we discuss some of the classical construction problems in Section 6–6.

EXERCISES

1. List five irrational constructible numbers.
2. Choose a unit segment and construct each of the numbers listed in Exercise 1.
3. Give the set of conjugates associated with each of the numbers listed in Exercise 1.
4. For each number listed in Exercise 1 give an irreducible equation with integral coefficients having the given number as a root.
5. Given a coordinate system, construct the roots of the following equations:

   (a) $3x - 5 = 0$,
   (b) $x^2 - 6x - 1 = 0$,
   (c) $2x^2 + 5x - 3 = 0$,
   (d) $x^4 + 3x^2 - 1 = 0$.

6. Find the distinct conjugates of each of the following:

$$1 + \sqrt{2}, \quad 2 - \sqrt{3 + \sqrt{2}}, \quad \sqrt{1 + \sqrt{2 - \sqrt{1 - \sqrt{2}}}}.$$

7. Find an irreducible equation with integral coefficients satisfied by each of the numbers given in Exercise 6.
8. Give three polynomial equations with integral coefficients that cannot be solved graphically using only straightedge and compasses.
9. It can be proved [15; 379] that a regular polygon of n sides may be constructed using only straightedge and compasses if and only if

$$n = 2^k p_1 p_2 \dots p_m,$$

where the p_i's are distinct prime numbers of the form $2^{2^t} + 1$. Indicate in the above form all values of $n \leqq 30$ such that a regular polygon of n sides may be constructed using straightedge and compasses. (One form of this result was first discovered by Gauss while he was still in his teens and greatly influenced his decision to devote his life to mathematics.)

6–6 Famous construction problems. There are three classical construction problems that challenged geometers for many centuries: the construction of a cube with volume double that of a given cube (the doubling or "duplication" of a cube), the construction of a square with area equal to that of a given circle (the squaring or "quadrature" of a circle), and the trisection of any given angle. These problems were known to the early Greeks and solutions are still being proposed. However, we may now use algebraic criteria (Sections 6–3 and 6–5) to establish that each of these problems is impossible under the classical restrictions. In Sections 6–7 and 6–8 we shall examine

a few nonclassical methods for solving the trisection problem and observe the manner in which each solution disregards the classical restrictions.

The construction of a cube with volume double that of a given cube. This problem is sometimes called the *Delian problem* since, according to tradition, it arose when the oracle at Delos advised the Athenians to double the size of the altar of Apollo. If the edge of the given cube is taken as the unit of length, the problem requires that a segment of length x be constructed, where $x^3 = 2$. The equation $x^3 - 2 = 0$ is irreducible in the ring of polynomials with integral coefficients, since by Theorem 4–9 it does not have a rational root. Then by Theorem 6–3 it does not have a constructible root. Thus it is not possible to construct $x = \sqrt[3]{2}$ and the Delian problem cannot be solved using only straightedge and compasses. The problem may be easily solved by nonclassical methods. For example, we could sketch the curve $y = x^3$ and find its intersection with the line $y = 2$.

The construction of a square with area equal to that of a given circle. If we take the radius of the given circle as the unit of length, this problem requires the construction of a root of the equation $x^2 = \pi$. This is possible only if the transcendental (Section 1–10) number π is constructible. We next observe that under the classical restrictions every constructible number is algebraic (Section 1–10) and therefore that no transcendental number may be constructed using only straightedge and compasses. Since every number constructible under the classical restrictions may be expressed in terms of the integers using rational operations and the extraction of square roots, every constructible number satisfies a polynomial equation with integral coefficients and accordingly is an algebraic number. Thus the transcendental number $\sqrt{\pi}$ is not constructible by classical methods, and the construction of a square with area equal to that of a given circle cannot be accomplished using only straightedge and compasses.

The proof that π is transcendental was first given by Lindemann in 1882, but nonclassical constructions of π have been known for many centuries [30; 55–80]. About 400 B.C. Hippias of Elis gave a nonclassical construction of a curve known as the quadratrix [55; 19–20] that may be used to obtain π and to trisect any angle. Briefly, given a quadrant of a circle OAB, as in Fig. 6–3, we consider a point Q moving at constant velocity along the arc AB and a point R moving at constant velocity along the radius OB in such a way that the two points start simultaneously from A and O respectively and arrive

simultaneously at B. At any time t_1 we may designate the positions of the points as R_1 and Q_1. The locus BPD of the intersection of the radius OQ_1 and the corresponding line through R_1 parallel to OA constitutes the *quadratrix*. The quadratrix may also be obtained by sliding the point R along the line OB at a constant rate and rolling the circle about O at a constant rate.

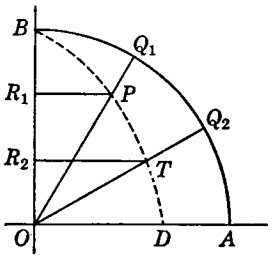


FIG. 6–3

The equation of the quadratrix may be obtained from the relations

$$y = ct, \quad \arctan y/x = kt, \tag{6–3}$$

where $P:(x, y)$ is a general point of the quadratrix and t indicates the time that the particles Q and R have been in motion. Let the unit of time be so chosen that R moves from O to B as t varies from 0 to 1. Then $\pi/2 = k$, and $y/x = \tan \pi t/2$ is an equation of the quadratrix. If we also take the unit of length in such a way that $OB = 1$, we have $1 = c$, whence $y = t$ and $x = t/\tan(\pi t/2)$. When $t = 0$, this relation may be used to obtain $OD = 2/\pi$ by using the methods for treating indeterminate forms that are discussed in the calculus. Thus $2/\pi$ and therefore π may be constructed using the quadratrix, straightedge, and compasses.

The trisection of any given angle. This problem is still very popular in the sense that solutions are still proposed nearly every year. These solutions are necessarily nonclassical constructions, although often not so intended. We shall first consider an algebraic proof that the trisection problem is impossible under the classical restrictions and then (Sections 6–7 and 6–8) consider a few nonclassical solutions.

The usual proof that there exists at least one angle that cannot be trisected using the classical methods makes use of a few trigonometric identities and an angle of 120°. Given an angle of 120°, we seek an angle of 40°. The construction of an angle of 40° is equivalent to the construction of a right triangle with hypotenuse unity and base cos 40°. Thus an angle of 120° can be trisected if and only if a segment of length cos 40° can be constructed. From $\cos 120° = -\frac{1}{2}$ and the trigonometric identity

$$4\cos^3 x - 3\cos x = \cos 3x,$$

we may obtain the relation

$$4\cos^3 40° - 3\cos 40° + \tfrac{1}{2} = 0.$$

Taking $y = \cos 40°$, we have

$$4y^3 - 3y + \tfrac{1}{2} = 0.$$

Multiplication by 2 and the substitution $z = 2y$ now gives

$$z^3 - 3z + 1 = 0.$$

This equation has no rational root (Theorem 4–9) and therefore is irreducible in the ring of polynomials with integral coefficients. Then by Theorem 6–3 it has no constructible root and $z = 2\cos 40°$ is not constructible, i.e., an angle of 120° cannot be trisected and the trisection problem is impossible under the classical restrictions.

The above proof that no root z is constructible using Theorems 4–9 and 6–3 may be rephrased in simpler language as follows. The only possible rational roots of the equation $z^3 - 3z + 1 = 0$ must be integers, since the leading coefficient is unity and any rational root must divide the constant term. Thus the only possible rational roots are $+1$ and -1. Since neither of these integers is a root, the equation has no rational roots. If the equation had a quadratic surd root such as $a + b\sqrt{2}$, it would also have the conjugate quadratic surd $a - b\sqrt{2}$ as a root and the remaining (third) root would be rational. Since any number constructible by classical methods may be expressed in terms of integers using a finite number of rational operations and the extraction of square roots, every nonrational constructible number is one of a set of an even number of conjugates (including itself) (Section 6–5). Thus any equation with integral coefficients having a constructible surd root must have an even number of surd roots. In particular, any cubic equation with integral coefficients having at least one constructible surd root must have exactly two surd roots and one rational root. Thus the fact that the above cubic equation has no rational root implies that it has no constructible root. As before, this implies that an angle of 120° cannot be trisected, whence not every angle can be trisected, and the trisection problem cannot be solved using only straightedge and compasses. In the next two sections we shall consider a few nonclassical methods for trisecting angles and see where they fail to observe the classical restrictions.

EXERCISES

1. Is it possible under the classical restrictions to divide an arbitrary angle into seven equal parts? Explain.
2. Give a classical construction of any sixteen points of the ellipse

$$x^2/4 + y^2 = 1.$$

3. Can the ellipse in Exercise 2 be completely graphed under the classical restrictions? Explain.

4. Give a nonclassical construction of the ellipse in Exercise 2.

5. State necessary and sufficient conditions on the coefficients of the equation $Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0$ in order that the complete graph of the equation may be drawn using only straightedge and compasses (Section 7–3).

6–7 Nonclassical geometric trisections. The classical restriction that there should be no marks upon the straightedge is disregarded in one of the simplest and oldest constructions of the trisections of angles. This construction is attributed to Archimedes. It requires only compasses and a straightedge with two marks on it. A ruler will serve very well.

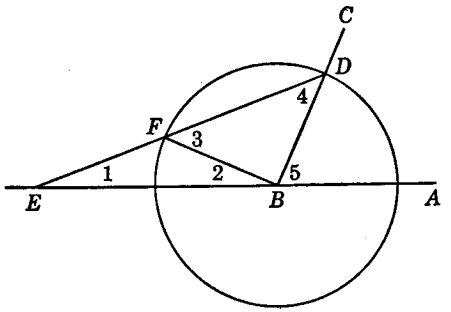


FIG. 6–4

Given any acute angle ABC (Fig. 6–4) construct a circle of radius r about the vertex B. Let D be the intersection of the circle and the side BC. Extend the side AB through B. On the straightedge mark off the length $r = BD$. Now keep the straightedge on D and slide one mark along AB extended until the other mark contacts the circle at some point F. Draw DF and let it intersect AB at E. Then $\angle BEF = \frac{1}{3}\angle ABC$. This fact may be proved as follows: Draw BF and designate the angles, as in Fig. 6–4. Then $EF = FB = BD = r$, $\angle 1 = \angle 2$, $\angle 3 = \angle 4$. Since $\angle 3$ is an exterior angle of triangle BEF and $\angle 5$ is an exterior angle of triangle BED, we have $\angle 3 = \angle 1 + \angle 2 = \angle 1 + \angle 1$, $\angle 5 = \angle 1 + \angle 4 = \angle 1 + \angle 3 = \angle 1 + \angle 1 + \angle 1$, and thus $\angle ABC = 3\angle BEF$. This is not a classical solution of the trisection problem, since marks are used on the straightedge.

Another nonclassical solution of the trisection problem was devised by Hippias of Elis about 400 B.C. (Section 6–6). This method involves the quadratrix (Fig. 6–3). From the relations (6–3), we have

$$\angle AOQ_2 : \angle AOQ_1 = OR_2 : OR_1,$$

whence angle AOQ_1 may be trisected by taking $OR_2 = \frac{1}{3}OR_1$, R_2T parallel to OA, and drawing OTQ_2. Then $\angle AOQ_2 = \frac{1}{3}\angle AOQ_1$.

This solution of the trisection problem does not satisfy the classical restrictions, since the quadratrix cannot be exactly drawn by classical methods.

Several proposed solutions of the trisection problem consist of a sequence of steps such that lines may be drawn as close as desired to the required trisecting lines. These solutions do not satisfy the classical restrictions, since one must be able to draw the required lines in a finite number of steps.

The popularity of the trisection problem is evident from the ever-increasing list of discoverers of methods for trisecting angles. The January 5, 1948 *Chicago Sun* carried an article entitled "Trisect an Angle? Simple . . . He Insists." The construction described in the article employs a circle with diameter equal to the width of the ruler used, places the upper left corner of the ruler along a certain line, and manipulates the ruler until the upper right corner meets another line. Although this construction does not use any marks on the ruler, it does use the fixed width of the ruler, contrary to the classical restrictions upon the problem (Section 6–1).

The frequent appearance of methods for trisecting angles emphasizes that mathematicians and teachers have not yet succeeded in making known the facts regarding the trisection problem, namely, that this problem can easily be solved by nonclassical methods but cannot be solved under the classical restrictions.

There are many other nonclassical constructions for solving the trisection problem [55]. Several of these involve the use of curves, such as the graph of

$$x^3 + xy^2 + ay^2 - 3ax^2 = 0,$$

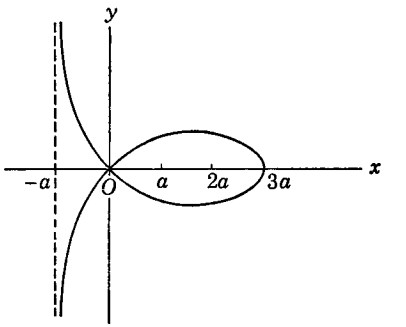


FIG. 6–5

(Fig. 6–5) that cannot be constructed using only straightedge and compasses. The curve in Fig. 6–5 is called the Trisectrix of Maclaurin and may be used to trisect any angle. Given a Trisectrix of Maclaurin and any angle ABC, draw a line m through the point $(2a, 0)$ making an angle with the positive x-axis equal to the given angle. Find the three intersections P_1, P_2, P_3 of the line m with the given curve. One of the lines P_iO, where O is the origin, makes an angle

with the positive x-axis equal to one-third of the given angle. It is not difficult to determine which of the three lines P_iO is to be used, since one may quickly compare three times each of the three angles obtained with the given angle. This method of trisecting angles is discussed in many analytic geometry texts. It does not observe the classical restrictions upon the problem, in that it employs a curve that cannot be constructed using only straightedge and compasses.

We have now considered several nonclassical constructions for solving the trisection problem and have observed the manner in which each method has failed to follow the classical restrictions upon the problem. In the next section we shall consider a few mechanical trisectors.

Exercises

1. Use Archimedes' construction and Exercise 22, Section 6–2 to find a method for trisecting angles of any size.

2. Draw angles of approximately 80°, 150°, 250°, 300°, and 750°. Trisect each of the angles just drawn.

3. Describe and illustrate Pappus' nonclassical solution of the trisection problem, using straightedge, compasses, and the construction of a hyperbola [55; 22–23].

6–8 Mechanical angle trisectors. There are several types of mechanical angle trisectors, varying in complexity from a coffee can cover with two sticks attached [4] to linkages (systems of bars connected by pin joints) in the construction of which very careful measurements must be made.

The simple angle trisector mentioned above may be made from any circular disk. It is a variation of the Tomahawk [55; 37]. Let the radius of the disk be r, the center O. Attach a stick OPQ of length $2r$ firmly to the disk. Then attach a second stick PT tangent to the disk at P (Fig. 6–6). This device may be used to trisect any given angle ABC by sliding TP through B until the disk is tangent to one side of the given angle, say at E, and Q is on the other side (Fig. 6–7). The right triangles QBP, OBP, and OBE are congruent, whence the lines BP and BO trisect the given angle.

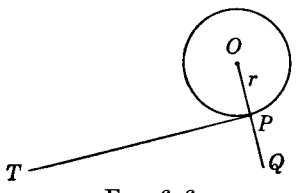


Fig. 6–6

Archimedes' construction (Section 6–7) provides a basis for a simple angle trisector made from four bars of wood [5]. Another type of angle trisector, Sylvester's Isoklinostat (Fig. 6–8), consists of four bars (OA, OC, OE, and OG) pivoted together at one end and joined by a system of shorter bars that keep the angles between the long bars equal (Exercise 1, Section 6–9). Since this type of mechanism could be used to divide an angle into any number of equal parts, it is called an "isoklinostat." The origin of Sylvester's basic idea for this linkage may be indicated by the following quotation from the title of the paper in which it was first published, "On a Lady's Fan, ..."

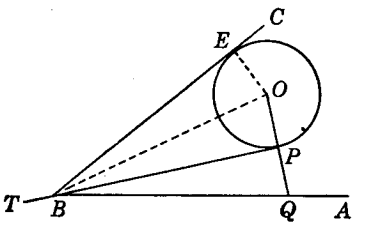


Fig. 6–7

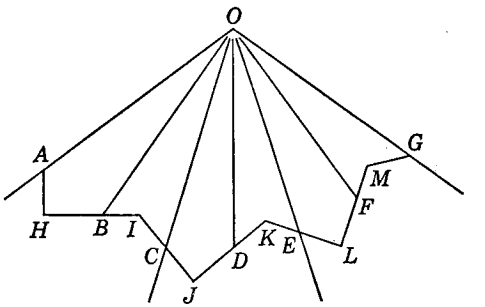


Fig. 6–8

Another angle trisector (linkage) was devised by a London barrister, Alfred Bray Kempe, in 1877. Kempe became interested in linkages after hearing Sylvester lecture on the subject. His angle trisector (Fig. 6–9) is based upon similar contraparallelograms (quadrilaterals with opposite sides equal and one pair of opposite sides crossing each other),

$$ABCD \sim ADEF \sim AFGH,$$

and can be used for any angle less than a complete revolution.

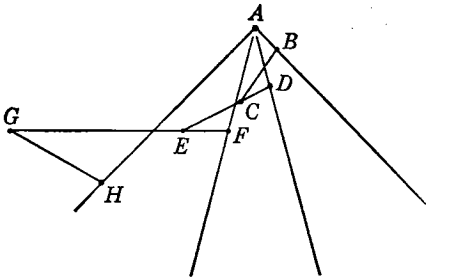


Fig. 6–9

We have now proved that the classical trisection problem is impossible and have seen that by using a protractor, linkage, marked ruler, suitable given curve, etc., the trisection problem may be easily solved by nonclassical methods. The next section contains a more general discussion of linkages.

6–9 Linkages. A linkage may be defined as a system of bars connected by pin joints to allow deformability without sliding motion. Between 1860 and 1895 considerable work was done with linkages. James Watt was dissatisfied with the mechanism used on the steam engine to change the straight-line motion of the piston into the circular motion of the wheels. This and more fundamental considerations led to a series of attempts to construct a theoretically straight line. Such a line (the inverse of a circle with respect to a point on it) was finally achieved by Peaucellier in 1864 and independently by Lipkin in 1871 [25]. Another solution was obtained by Bricard in 1895. The rapid development and popularity of the subject is illustrated by the fact that six of the twenty-six papers presented at the annual General Meeting of the London Mathematical Society on November 11, 1875 were on linkages. Probably the peak in this development occurred in 1876 when Alfred Bray Kempe presented to the London Mathematical Society a paper "On a General Method of describing Plane Curves of the nth degree by Linkwork," showing how to construct a linkage to trace any plane curve $f(x, y) = 0$ of the nth degree. Thus, theoretically, the graph of any plane polynomial curve may be drawn by linkages. However, a glance at some of the diagrams of linkages for higher degree curves [45] will indicate that the linkages became too complicated to be useful in this respect. It has also been proved [2; 52] that no transcendental curve may be drawn using linkages. Recently, except for a few articles recommending their use as visual aids [36], [41] or material of general interest [25], [55], very little has been said about linkages.

The recent lack of publicity on linkages does not, however, indicate that they are obsolete. Actually, many applications for linkages have been found, and they are used a great deal. A description of the linkage computers developed at the Radiation Laboratory during World War II fills a good-sized book. The pantograph (Exercise 4, Fig. 6–10) is used for copying figures (similar figures) and for plotting equipotential points in an electrical field. The mechanism whereby public officials sign many checks at a time is a form of linkage. Nearly all machinery contains linkages — either obvious or disguised — as may be readily verified by consulting a textbook

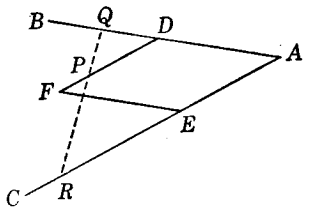


FIG. 6–10

on mechanisms. From the teacher's viewpoint, linkages can be very useful in showing students that geometry is not only static but also dynamic [36], [41].

Linkages give rise to only one of several methods of nonclassical construction. We shall conclude this chapter with a brief general discussion of classical and nonclassical constructions.

EXERCISES

1. Given that the points A, B, C, D, E, F, G in Fig. 6–8 lie on a circle with center O, show that $\angle AOC = \angle COE = \angle EOG$ if $AH = HB = CJ = JD = EL = LF$ and $BI = IC = DK = KE = FM = MG$, where AH, HI, IJ, JK, LM, and MG are each single bars.

2. Design a linkage that will divide any given angle into five equal parts.

3. Give two examples of the use of linkages to illustrate dynamic (as contrasted with static) mathematical properties in dealing with geometrical figures.

4. A pantograph may be constructed using two long bars AB, AC, and two short bars $FD = AE$, joined as in Fig. 6–10 such that $AD = FE$. Assume that the point Q on AB is held fixed and show that as P traces any figure, R traces a similar figure, where P, Q, R are collinear, i.e., show that the ratio QP/QR is constant.

6–10 Summary. The three famous construction problems (Section 6–6) were formulated [2; 20] as early as the fifth century B.C. During that same century Hippias of Elis gave a nonclassical solution of the trisection problem using the quadratrix. A few years later this same curve was used to solve the problem of squaring a circle. Also the conchoid was used in the trisection of an angle and the duplication of a cube. Thus by the end of the third century B.C. all three of these famous problems could be solved by nonclassical methods, the Greek mathematicians were reaching the peak of their development, the foundations of our plane geometry were being developed by Euclid, and geometry as a science was beginning to emerge. However, mathematicians were to wait over two thousand years for sufficient development of algebra and geometry to prove that the three famous construction problems could not be solved using only straightedge and compasses.

The early Greek mathematicians felt that every construction of elementary geometry could be made using only straightedge and compasses. Thus these classical constructions have played an important role in the development of geometry. In one sense they are

the forerunners of projective geometry. In the tenth century, an Arabian mathematician considered constructions with straightedge and compasses with a fixed opening. Much later (nineteenth century) Poncelet and also Steiner proved that the constructions with straightedge and compasses are exactly equivalent to the constructions with straightedge and a fixed circle [2; 30]. In 1672 Georg Mohr and in about 1800 Lorenzo Mascheroni proved that only compasses were needed to obtain any point that could be constructed using straightedge and compasses. Analytic geometry was now being developed, and by the nineteenth century Gauss was able to obtain analytic criteria for the constructibility of regular polygons. Others discovered that the roots of the equations arising from the three famous construction problems could not be constructed, and thus that these problems could not be solved using only straightedge and compasses.

Linkages developed rapidly during the last part of the nineteenth century. It was soon shown (Section 6–9) that every algebraic curve, and in particular a straight line, could be traced using linkages. Beginning about this time and continuing into the present century Felix Klein contributed a great deal to our understanding of nearly all the above material through his famous lectures.

In the present chapter we have considered the classical constructions with straightedge and compasses in detail (Sections 6–1 to 6–6). We have also mentioned a few methods used in nonclassical constructions (Sections 6–7 to 6–9). Marks on a ruler or suitable given curves may be used to trisect arbitrary given angles. Linkages may be used for a great variety of constructions. A regular polygon of any finite number of sides may be inscribed in a given circle, using a protractor. Some devices such as those just mentioned enable one to solve problems that cannot be solved by classical methods. Other common devices, such as dividers, parallel ruler, fixed square, fixed circle, merely permit short-cuts in some classical construction problems without making it possible to solve any additional problems. In general, any construction problem is solvable if there are no restrictions upon the methods that may be used; some construction problems are not solvable (Theorem 6–1) when only straightedge and compasses may be used. Accordingly, we have used the results obtained by Gauss, Klein, and others, considered many common classical constructions, developed an algebraic criterion for determining whether or not a given construction may be performed using only straightedge and compasses, applied this criterion to several classical problems, considered a few nonclassical constructions, and briefly mentioned some modern applications of these methods in teaching and industry.

We have emphasized the famous classical construction problems (Section 6–6) and especially the trisection problem. This problem has been used to illustrate the application of algebraic methods to the solution of classical construction problems. We have seen that algebraic theories may be used to show that all attempts to obtain a classical solution are necessarily in vain. This application of algebraic theories to geometric constructions provides one example of the interdependence of algebra and geometry. The remaining chapter provides another example of this interdependence through a consideration of graphical representations of certain common algebraic functions.

Exercises

1. Construct regular polygons of 6, 7, 8, and 9 sides.
2. Specify which of the following sets of numbers contain only constructible numbers under the classical restrictions: integers, rational numbers, algebraic numbers, real numbers.
3. Repeat Exercise 2 for numbers constructible using linkages.
4. Describe three modern machines in which linkages are used.
5. Describe the linkages in three instruments or common machines that you have used.
6. Discuss Gauss' contribution to the study of constructions.

CHAPTER 7

GRAPHICAL REPRESENTATIONS

Integers, rational numbers, constructible numbers (Section 6–4), real numbers (Section 1–12) may be represented as points on a line in Euclidean geometry. Complex numbers may be considered as ordered pairs of real numbers (Section 1–15) and represented as points in a plane in Euclidean geometry. Any single-valued function of a real variable x may be used to obtain ordered pairs of numbers and may be represented graphically. These representations underlie our present consideration of algebraic concepts in terms of geometric concepts. Throughout this chapter we shall be primarily concerned with the graphs in (real) Euclidean spaces of algebraic functions with real coefficients. We shall consider the graphs of several types of functions, various methods of constructing the graphs, and several applications of graphs and graphical methods.

7–1 Euclidean and complex spaces. The one-to-one correspondence (Cantor-Dedekind Axiom, Section 1–12) between the set of points on a line in Euclidean geometry and the set of real numbers provided a basis for a geometrical interpretation of the four rational operations in Chapter 6. These concepts can be used to develop an isomorphism (Section 1–8) between the set of points on a line in Euclidean geometry and the set of real numbers. There also exist isomorphisms between pairs of real numbers and the set of points on a Euclidean plane, triples of real numbers and the set of points in Euclidean 3-space, and, in general, n-tuples of real numbers and the set of points in Euclidean n-space.

On the basis of the above isomorphisms, any point on a line in Euclidean geometry may be uniquely identified by one real number (coordinate), any point on a Euclidean plane may be identified by two real coordinates, any point on a Euclidean 3-space by three real coordinates, . . . , any point on a Euclidean n-space by n real coordinates. Thus in Euclidean geometry we speak of a line as a one-dimensional space, speak of a plane as a two-dimensional space, and, in general, discuss n-dimensional Euclidean spaces for any positive integer n.

It is also often convenient to refer to a set of points that may be made isomorphic with the set of complex numbers as a one-dimensional complex space. Since a complex number may be considered as an ordered pair of real numbers, a one-dimensional complex space is isomorphic with a Euclidean plane. Similarly, a space with two complex coordinates is isomorphic with a Euclidean 4-space. In general, a complex n-space is isomorphic with a Euclidean $2n$-space.

We have not explicitly considered distance or metric relations in the above brief descriptions of Euclidean and complex spaces. In a complete treatment, the distance relations would be involved in the establishment of the isomorphisms mentioned above.

Any function of n variables that vanishes for one or more sets of real values of the variables (real n-tuples) has a graph in a (real) Euclidean n-dimensional space, i.e., the set of all points with coordinates (n-tuples) that make the function zero. The function $x^2 + y^2 - 1$ has the unit circle about the origin in the Euclidean xy-plane as its graph. The function $x^2 + y^2 + 1$ has no real zeros and is said to have a *vacuous graph* in the xy-plane. Thus a polynomial function of n variables may or may not have a nonvacuous graph in a Euclidean n-space (Section 7–2).

The situation is quite different in a space with complex coordinates. Every polynomial $f(x)$ with complex coefficients and of positive degree has a graph in a space with one complex coordinate (Theorem 4–3). For example, $x^2 - 1$ has as its graph the points $+1$ and -1, whether the points are considered to be on the real line or a space with one complex coordinate such as that discussed in Section 1–16; $x^2 + 1$ has a vacuous graph on the real line but has the points i and $-i$ as its graph in a space with one complex coordinate. This property of polynomials $f(x)$ may be extended (Exercises 4, 5, and 6) to show that every algebraic function of n variables has a nonvacuous graph in a space with n complex coordinates.

Exercises

1. Give four functions of three variables (a) that have a nonvacuous graph in E_3, (b) that have a vacuous graph in E_3.
2. Give three functions of n variables (a) that have a nonvacuous graph in E_n, (b) that have a vacuous graph in E_n.
3. Prove that any polynomial in one variable with complex coefficients has a nonvacuous graph in a space with one complex coordinate.

4. Prove that any polynomial in n variables with complex coefficients has a nonvacuous graph in a space with n complex coordinates.

5. Prove that any algebraic function in one variable with complex coefficients has a nonvacuous graph in a space with one complex coordinate.

6. Prove that any algebraic function of n variables with complex coefficients has a nonvacuous graph in a space with n complex coordinates.

7–2 Polynomials. Any linear polynomial in n variables with real coefficients always has a real graph in a Euclidean n-dimensional space E_n. This graph is a point in E_1, a line in E_2, a plane in E_3, and, in general, a hyperplane in E_n. Thus a real linear function of n variables has as its graph an E_{n-1} in E_n for all positive values of n. Each of these graphs, E_{n-1} spaces, divides the corresponding E_n into two regions, on one of which the function is positive and on the other, negative. The graphs are called linear subspaces and play an important role in many advanced mathematical theories.

Given any linear equation in n variables ($n > 1$) with real coefficients, arbitrarily many real n-tuples of numbers for which the equation is satisfied may be found by rational operations. Thus the coordinates of arbitrarily many points on the graph may be found by rational operations. The graph is completely determined by n linearly independent (Section 5–13) n-tuples (points). For example, a plane is completely determined by three points that are not on the same line (Section 5–14) and, in general, an E_{n-1} is completely determined by n points that are not on the same E_{n-2}. When the coefficients of the n variables and the constant term are all real and different from zero, the n real, distinct points at which the graph intersects the coordinate axes completely determine the graph.

Any polynomial of degree n in one variable with complex coefficients has n complex zeros (Theorem 4–2). A polynomial of degree n with real coefficients has n complex zeros but may or may not have real zeros, as the examples $x^2 - 1$, $x^2 + 1$ given above illustrate. Thus a real polynomial equation in one variable may or may not have a nonvacuous graph on the real line. By exactly the same reasoning, a real polynomial in two variables may or may not have real zeros corresponding to a given value of one of the variables. For example, $x^2 + y^2 - 25$ has zeros $+3$ and -3 when $x = 4$, but no real zeros when $x = 6$. If a polynomial such as $x^2 + y^2 + 1$ has no real zeros for every real value of x, it has a vacuous graph in the Euclidean plane. In general, if a polynomial in n variables has no real zeros for every real set of values for $n - 1$ of the variables, it has a vacuous graph in Euclidean n-space.

The graph of a polynomial divides the space into regions on which the polynomial has constant sign, since any polynomial is a continuous function of its variables. A polynomial in one variable changes sign if and only if the variable passes through a zero of odd multiplicity (Section 4–13). For two or more variables the path of the general point $(x_1, x_2, \ldots, x_n)$ in passing through a point on the graph of $f(x_1, x_2, \ldots, x_n)$ must be considered. For example, the polynomial $x^2 + y^2 - 1$ changes its sign at $(1, 0)$ as the point (x, y) traverses the line $y = 0$, but does not change its sign at $(1, 0)$ as the point (x, y) traverses the line $x = 1$. In general, the multiplicity m of an intersection point P of a curve C with the graph of a polynomial in n variables may be so defined that as the point $(x_1, x_2, \ldots, x_n)$ traverses the curve C, the polynomial changes sign at P if and only if m is odd. We shall define only the multiplicity of the intersection of a line and a polynomial curve in a plane.

Suppose a given polynomial curve $f(x', y')$ and a given line intersect at a point $P:(s, t)$ with a multiplicity k, where k is to be determined. By the change of variables (translation, Section 5–15) $x = x' - s$, $y = y' - t$, the curve $f(x + s, y + t) = g(x, y)$ and the given line expressed in the new coordinates intersect at the new origin with the same multiplicity k as the curve $f(x', y')$ intersected the given line at P. The equation of the line now has either the form $y = mx$ or the form $x = 0$. After substituting in these two cases, we consider respectively the polynomials $g_1(x) = g(x, mx)$ and $g_2(y) = g(0, y)$. The value of k is then determined by the fact that the terms of lowest degree in $g_1(x)$ or, if the line is $x = 0$, $g_2(y)$ have degree k. According to this definition, a line and a curve that do not both pass through the point P are said to have an *intersection of multiplicity zero* at P.

The graph of a polynomial in n variables divides a Euclidean n-dimensional space into a finite number of regions, on each of which the polynomial has constant sign. The general problem of determining these regions [the solutions of $0 < f(x_1, x_2, \ldots, x_n)$ and $0 < -f(x_1, x_2, \ldots, x_n)$] and their boundaries [the solutions of $0 = f(x_1, x_2, \ldots, x_n)$] has not been completely solved. However, considerable work has been done for polynomials in two or three variables. In particular, we shall now consider real polynomials of degree two in two variables (Section 7–3) and real polynomials of degree two in three variables (Section 7–4).

EXERCISES

1. Write a linear equation in n real variables and find a set of n points that determine its graph when n is (a) 2, (b) 3, (c) 4, (d) 5, (e) 6, (f) 10.

2. Find the multiplicity of the intersection at the origin of the line $y = 0$ with each of the following curves:

(a) $y = x^2$, (d) $x^3 + 3x^2y + x^2 = 0$,
(b) $x = y^3$, (e) $x^2 + y^2 = 1$,
(c) $y^4 = x^2$, (f) $x = 0$.

3. Find the multiplicity of the intersection at the origin of each of the following lines with each of the curves in Exercise 2: (a) $x = 0$, (b) $y = x$.

4. Find the multiplicity of the intersection at (2, 1) of the line $y = 1$ with each of the following curves:

(a) $x + y = 3$, (d) $(x - 2)^2 + y^2 = 1$,
(b) $x^2 + y^2 = 5$, (e) $x^3 = 8y$,
(c) $x = 2y^2$, (f) $x^2 - 3y^2 = 1$.

5. Repeat Exercise 4, using each of the following lines: (a) $x = 2$, (b) $x = 2y$.

7–3 Conic sections. The general real quadratic equation in two variables has the form

$$Ax^2 + Bxy + Cy^2 + Dx + Ey + F = 0, \tag{7–1}$$

where the coefficients are assumed to be real and $A^2 + B^2 + C^2 \neq 0$. The graphs of equations of the form (7–1) are called *conic sections*, since for every set of real coefficients for which (7–1) has a non-vacuous graph, the graph may be obtained as the intersection of a plane and a right circular cone (possibly degenerate). In this section we shall discuss briefly the development of the hyperbola, parabola, ellipse, and circle as the intersections of planes with a right circular cone, i.e., as plane sections of a right circular cone. We shall also mention several of the properties of these conic sections as a review of analytic geometry.

A circle may be defined as the locus of points in a plane equidistant from a given fixed point Q of the plane. When there is a coordinate system (x, y) and a distance relation in the plane, the circle is the graph of a polynomial $(x - h)^2 + (y - k)^2 - r^2$, where its center Q has coordinates (h, k) and the distance is $r \geqq 0$. A circle with $r = 0$ is called a *point circle* and is classified as a degenerate form of the circle.

Consider a nondegenerate real circle with center Q and let $P \neq Q$ be an arbitrary real fixed point on the line that passes through Q and is perpendicular to the plane of the circle. The locus of points on the set of lines joining P to points of the circle is called a *right circular cone* (Fig. 7–1). The cone has two *nappes* meeting at P. The locus of points on the set of lines perpendicular to the plane of the circle and through points of the circle is called a *right circular cylinder* (Fig. 7–2) and may be considered as a limiting case of the cone as the distance PQ increases without bound.

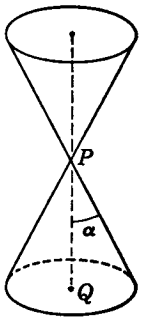


FIG. 7–1

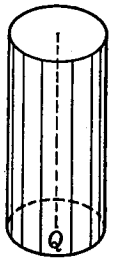


FIG. 7–2

Let π be an arbitrary real plane and consider the right circular cone generated by lines through the fixed point P and a given nondegenerate circle with center Q as above. Let the constant angle between the generating lines and PQ be α. A graph of the quadratic equation (7–1) is said to be degenerate if and only if it is obtained by passing the plane π through P. This condition may also be expressed algebraically [18; 215–219] as follows: The graph is degenerate if and only if $\Delta = 0$, where

$$\Delta = \begin{vmatrix} 2A & B & D \\ B & 2C & E \\ D & E & 2F \end{vmatrix}.$$

A nondegenerate graph of (7–1) is a hyperbola, parabola, or ellipse according as $B^2 - 4AC >, =, < 0$. Geometrically, these three cases, respectively, arise according as the smallest angle θ between the plane π and the line PQ is $<, =, > \alpha$ (Fig. 7–3). (The normal to the plane makes an angle with PQ equal to the complement of θ.) For example, consider the cone $3x^2 + 3y^2 - z^2 = 0$, with $\alpha = 30°$. For any real number k the plane $z = k$, with θ equal to a right angle, intersects the cone in a circle $3x^2 + 3y^2 - k^2 = 0$; the plane $z = x + k$ with $\theta = 45° > \alpha$ intersects the cone in an ellipse $2x^2 + 3y^2 - 2xk - k^2 = 0$; the plane $z = x\sqrt{3} + k$ with $\theta = 30° = \alpha$ intersects the conic in a parabola $3y^2 - 2xk\sqrt{3} - k^2 = 0$; and the plane $z = 2x + k$ with $\theta < \alpha$ intersects the cone in a hyperbola $3y^2 - x^2 - 4xk - k^2 = 0$.

The degenerate conics may be identified algebraically or geometrically. From a geometric point of view (Exercise 1), a hyperbola

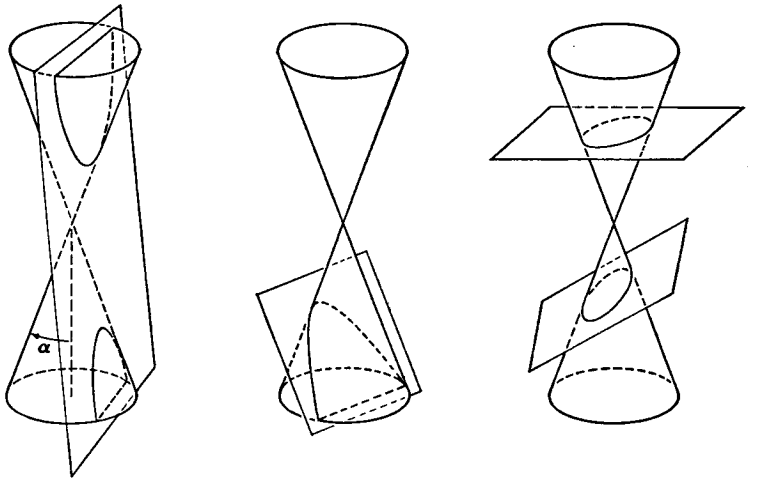


FIG. 7-3

may degenerate into two intersecting lines, a parabola into two coincident lines or two parallel lines (using a right circular cylinder), and an ellipse into a point. The ellipse becomes a circle when θ is a right angle.

If the coordinate axes are rotated (Section 5–15) through an angle ψ, where $\tan 2\psi = B/(A - C)$ if $A \neq C$ and $\psi = 45°$ if $A = C$, it is shown (Exercise 6) in most analytic geometry texts that the equation (7–1) takes on the form

$$A'x^2 + C'y^2 + D'x + E'y + F' = 0. \tag{7–2}$$

It can also be shown (Exercise 7) that the numbers $A + C$, $B^2 - 4AC$, and Δ are unchanged by a rotation or translation of the coordinate axes [11; 100]. Thus $B^2 - 4AC = -4A'C'$ and the graph, possibly degenerate, of (7–2) is a hyperbola, parabola, or ellipse according as $A'C' <, =, > 0$. Since A' and C' cannot both be zero in the quadratic equation (7–2), the general equation of a nondegenerate parabola may be written in one of the forms

$$(y - k)^2 = 2p(x - h) \quad \text{or} \quad (x - h)^2 = 2p(y - k). \tag{7–3}$$

Similarly, if $A'C' \neq 0$, let $h = -D'/2A'$ and $k = -E'/2C'$. Then a nondegenerate ellipse has an equation of the form

$$\frac{(x - h)^2}{a^2} + \frac{(y - k)^2}{b^2} = 1, \quad 0 < a, \quad 0 < b \tag{7–4}$$

and a nondegenerate hyperbola has an equation of the form

$$\frac{(x - h)^2}{a^2} - \frac{(y - k)^2}{b^2} = 1 \quad \text{or} \quad \frac{(y - k)^2}{a^2} - \frac{(x - h)^2}{b^2} = 1. \tag{7–5}$$

The first parabola in (7–3) has axis $y = k$, vertex (h, k), and passes through the points $(h + p/2, k \pm p)$. It may be defined in the plane as the locus of points equidistant from the line $x = h - p/2$ (called the *directrix*) and the point $(h + p/2, k)$ (called the *focus*).

The ellipse (7–4) has center (h, k) and, assuming $a^2 > b^2$, the ends of its major axis are at $(h \pm a, k)$, the ends of its minor axis are at $(h, k \pm b)$, and its foci are at $(h \pm \sqrt{a^2 - b^2}, k)$. If $a^2 = b^2$, it is a circle. The ellipse may be defined in the plane as the locus of points P such that $PF_1 + PF_2 = 2a$, where F_1 and F_2 are the foci.

The first hyperbola in (7–5) has center (h, k), ends of its major axis at $(h \pm a, k)$, foci at $(h \pm \sqrt{a^2 + b^2}, k)$, and asymptotes $b(x - h) = \pm a(y - k)$. It may be defined in the plane as the locus of points P such that $PF_1 - PF_2 = \pm 2a$.

Thus the real graph (when such exists) of the general quadratic equation (7–1) may be obtained as the section of a right circular cone (possibly degenerate) by a plane and is called a conic section. The form of the graph may be specified in terms of the quantity $B^2 - 4AC$ and the rank (Section 5–10) of the determinant Δ. The definitions of hyperbola, parabola, and ellipse as loci on a plane can be proved to be equivalent to the definitions as sections of a right circular cone. A very readable treatment of conic sections may be found in [38; 102–138], a more complete treatment in [18; 171–236], a history of conic sections and quadric surfaces in [11].

EXERCISES

1. Draw figures illustrating how each of the following nonvacuous degenerate conics may be obtained as a plane section of a right circular cone or a right circular cylinder: (a) two intersecting lines, (b) two coincident lines, (c) two distinct parallel lines, (d) a point.

2. Graph the following conic sections:

(a) $x^2 + y^2 = 25$,
(b) $9x^2 + 4y^2 = 36$,
(c) $9x^2 - 4y^2 = 36$,
(d) $x^2 = y + 2$,
(e) $x^2 - 2x = y$,
(f) $x = y^2 - 2y + 5$,
(g) $y = x^2 - 6x + 7$.

3. Identify the graphs of the following equations:
   (a) $x^2 - 2y^2 + 3x + y = 5$,
   (b) $x^2 - 2xy + y^2 - 2x + y + 7 = 0$,
   (c) $xy = 12$,
   (d) $x^2 + 2xy + y^2 + 2x + 2y + 1 = 0$,
   (e) $3x^2 + 2xy - y^2 + 5x - 2y + 1 = 0$,
   (f) $x^2 + 2xy + y^2 + x + y - 6 = 0$,
   (g) $2x^2 - xy + 3y^2 - 4x + 6y = 0$.
4. Rewrite each of the equations in Exercise 3 in the form (7–2).
5. Graph the conic sections in Exercise 3.
6. Derive the equation (7–2) from (7–1) by considering the effect on (7–1) of a rotation (Section 5–15), showing how to choose an angle of rotation such that the xy term drops out, and expressing the new coefficients in terms of the old coefficients and the angle selected.
7. Show that each of the following expressions is invariant under the rotation used in Exercise 6: (a) $A + C$, (b) $B^2 - 4AC$, (c) Δ.
8. Find the rank (Section 5–10) of the determinant Δ corresponding to each equation in Exercise 3. Consider the general significance of the rank of Δ.

7–4 Quadric surfaces. The graph of a quadratic equation $f(x, y, z) = 0$ in three variables with real coefficients is called a *quadric surface*. If one of the variables, say z, is missing in the equation $f(x, y, z) = 0$, then $f(x, y, z)$ may be written as $f(x, y)$ and the graph in three dimensions (quadric surface) of the quadratic equation $f(x, y) = 0$ intersects every plane $z = c$ in a conic section (Section 7–3) congruent to the graph of $f(x, y)$ in the xy-plane. Throughout this chapter we shall speak interchangeably of the graph of $f(x, y) = 0$ and the graph of $f(x, y)$. This terminology is analogous to our previous consideration of the roots of a polynomial equation $f(x) = 0$ and the zeros of a polynomial $f(x)$. The graph of $f(x, y)$ in three dimensions consists of all points on lines parallel to the z-axis and through points of the graph of $f(x, y)$ in the xy-plane. The graph in three dimensions is a special case of a cylinder (not necessarily circular). Formally, a *cylinder* may be defined as a surface consisting of all points on lines that are parallel to a fixed line and which pass through points of a fixed curve in a plane that is not parallel to the fixed line. Thus the graph of any real quadratic equation $f(x, y, z) = 0$ having one of the variables missing may be considered as a cylinder having a coordinate axis as the fixed line and a conic section as the fixed curve in the coordinate plane that does not contain the fixed line. Any plane parallel to the plane of the fixed curve intersects the cylinder in a conic section congruent (under a translation) to the fixed curve.

Suppose $f(x, y, z) = 0$ is any real quadratic equation in the three variables x, y, z, and $mx + ny + rz + d = 0$ has any real plane as its graph (Section 7–2). Then at least one of the coefficients m, n, r is different from zero, and there exists a rotation in space such that under the new coordinate system the above plane has equation $z = c$ and the quadric surface has equation $g(x, y, z) = 0$, where $g(x, y, z)$ is a real quadratic polynomial. The intersection of the plane and the quadric surface is now on the cylinder $g(x, y, c) = 0$ and, since $g(x, y, c)$ is a real quadratic polynomial in x and y, this intersection is a conic section. We thus find, by rotating the coordinate system such that the new z-axis is perpendicular to the given plane, that any plane section of a quadric surface is a conic section.

The graph of any quadric surface may be obtained by considering the plane sections (conic sections) parallel to the coordinate planes. For example, $b^2x^2 + a^2y^2 = a^2b^2z$, where $ab \neq 0$ (Fig. 7–4), has an elliptic section in the plane $z = c$ for all positive values of c, a point for $c = 0$, and no real graph when c is negative. It has a parabolic section for all real values of d when $x = d$ or $y = d$, and is called an elliptic paraboloid.

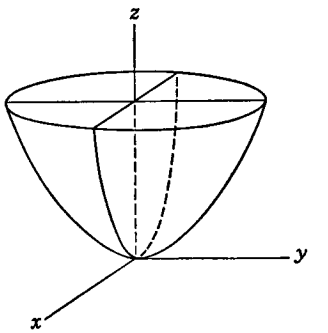


FIG. 7–4

The problem of obtaining the quadric surface from its plane sections parallel to the coordinate planes is exactly analogous to that of obtaining a conic section from its line sections parallel to the coordinate axes. For example, the parabola $y = x^2$ intersects every line $x = a$ in a single point and intersects the line $y = b$ in two distinct points when b is positive, two coincident points when $b = 0$, and two imaginary points (vacuous graph in the Euclidean plane) when b is negative. The graph of the parabola can be visualized from these intersections, using the fact that the graph is continuous (Sections 3–12 and 3–13). This method for determining graphs may be used for contour lines representing points of the same elevation on maps, isothermal lines representing points of the same temperature, and many other applications of level curves as discussed in some calculus textbooks. Thus, given any polynomial $f(x, y)$, the problem of visu-

alizing the surface $z = f(x, y)$ from the plane sections in which $z = c$ is the same as that of visualizing the topography of a landscape from the contour lines on a map.

The method of determining graphs by means of sections may also be used to visualize graphs in four dimensions. In this case the sections are taken by three-spaces parallel to the coordinate three-spaces. For example, $x^2 + y^2 = 1$ may be graphed as a unit circle in the xy-plane or as a cylinder in three-space such that every section by a plane $z = c$ is a unit circle. Similarly, $x^2 + y^2 + z^2 = 1$ may be graphed as a unit sphere in three-space or as a cylinder in four-space such that every section by a three-space $w = c$ is a unit sphere. The limitations of this method in four-space, five-space, etc. lie in the fact that we are accustomed to three-dimensional space and in the powers of visualization developed by the individual using the method (Exercises 9 to 14).

Some quadric surfaces, such as the hyperboloid of one sheet,

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1$$

(Fig. 7–5), contain straight lines and are called *ruled surfaces*. For example, consider the two pairs of planes

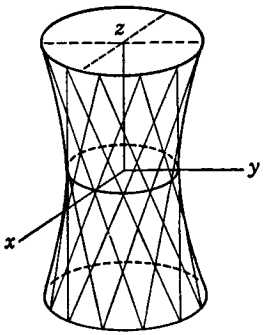


FIG. 7–5

$$\frac{x}{a} - \frac{z}{c} - k\left(1 + \frac{y}{b}\right) = 0,$$

$$1 - \frac{y}{b} - k\left(\frac{x}{a} + \frac{z}{c}\right) = 0$$

and

$$\frac{x}{a} + \frac{z}{c} - m\left(1 - \frac{y}{b}\right) = 0,$$

$$1 + \frac{y}{b} - m\left(\frac{x}{a} - \frac{z}{c}\right) = 0.$$

For every real value of k the intersection of the first pair of planes is a line on the quadric surface,

$$\frac{x^2}{a^2} - \frac{z^2}{c^2} = 1 - \frac{y^2}{b^2},$$

obtained by eliminating k between the equations of the two planes. Thus one obtains a line on the quadric surface for each real value of k. Similarly, one obtains a second set of lines on the quadric

surface by considering real values of m in the equations of the second pair of planes. When complex coordinates and coefficients are used, every central conic (hyperbola or ellipse) has a ruled graph. Cones and cylinders are common examples of ruled surfaces in three-dimensional real space. Real quadric surfaces may be classified as ellipsoids, elliptic paraboloids, etc. in terms of the coefficients of the second degree terms and the number of lines of the surface through each point of the surface. Most analytic geometry texts consider a few special quadric surfaces. Our purpose in this section has been to indicate the existence of a rather complete theory and classification of quadric surfaces analogous to those for conic sections. Further details may be found in [11], [16], and [18].

EXERCISES

1. Give a general real quadratic equation in three variables.
2. Give the equations of five cylinders in three-space having different types of fixed curves.
3. Graph the fixed curves of the cylinders given in Exercise 2.
4. Graph the cylinders given in Exercise 2.
5. Discuss the sections by all planes $z = c$ of each of the following quadric surfaces:

(a) $4x^2 + 9y^2 + 4z^2 = 36$,
(b) $x^2 = 2y$,
(c) $x^2 - y^2 = 1$,
(d) $x^2 + y^2 = z^2$,
(e) $x^2 - z^2 = 0$,
(f) $x^2 - 4y^2 = 8z$.

6. Repeat Exercise 5 for all planes $x = a$ and $y = b$.
7. Graph the quadric surfaces in Exercise 5.
8. Give the names of the quadric surfaces in Exercise 5.
9. Discuss the graph of $x - 1$ in the following spaces: (a) x-axis, (b) xy-plane, (c) xyz three-space, (d) $xyzw$ four-space, (e) $xyzwu$ five-space.
10. Repeat Exercise 9 for $x^2 - 1$.
11. Discuss the graphs of $x^2 + y^2 = 4$ in the spaces in Exercise 9(b)–(e).
12. Repeat Exercise 11 for $y = x^2$.
13. Discuss the graphs in the $xyzw$ four-space of the polynomial equations in Exercise 5.
14. Discuss the graphs in four-space of
    (a) $x^2 + y^2 + z^2 + w^2 = 1$,
    (b) $x + y + z + w = 1$,
    (c) $x^2 + y^2 = z^2 + w^2$.
15. How many lines on a cone (Fig. 7–1) pass through each point of the surface?

16. How many lines are there on each point of a cylinder in three-space having a nondegenerate conic section as its fixed curve?

17. Given a general real quadratic equation in three variables (Exercise 1), write down its determinant Δ analogous to that for the general conic section in Section 7–3.

18. Find the rank of Δ for each of the quadric surfaces in Exercises 2 and 5. Consider the general significance of the rank of Δ.

7–5 Higher plane curves. The graphs in the xy-plane of polynomials $f(x, y)$ of degree greater than two are called *higher plane curves*. These graphs have been completely classified when $f(x, y)$ has degree three or four, i.e., for cubic and quartic curves. Many other curves have been extensively studied. In this section we shall define a singular point and, in particular, a double point. Then we shall classify double points and finally classify cubic plane curves in terms of their double points. A few general properties of higher plane curves will be mentioned.

A point P of a curve such that every line through P intersects the curve with a multiplicity (Section 7–2) at least two at P is called a *singular point* of the curve. If some line through a singular point P intersects the curve with multiplicity two at P, then P is a *double point*. Any line is either entirely on (i.e., a component of) a curve of degree n or intersects the curve in at most n points. This can be proved using the fact that if an equation $f(x, y) = 0$ of degree n and the equation of the line are solved simultaneously, then the resulting equation in one variable is either identically zero or of degree at most n. Similarly, two curves of degree m and n either have a component in common or intersect in at most mn points. Using such arguments, it can be shown that a curve of degree n can have at most $\frac{1}{2}(n-1)(n-2)$ double points [23; 41–42]. In particular, a cubic curve has at most one double point (Exercise 1).

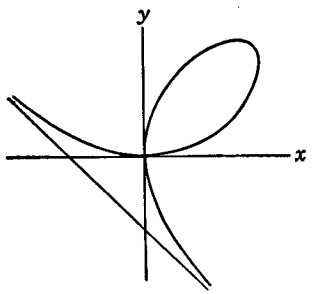


FIG. 7–6

At an ordinary (nonsingular) point, a curve of degree n has a unique tangent; at a double point it has two tangents; and, in general, at a singular point P it has k tangents when every line through P intersects the curve with multiplicity at least k at P and some line intersects the curve with multiplicity exactly k at P.

Double points are classified as *nodes* when the tangents are distinct, *cusps* when they coincide. When the tangents are conjugate imaginary lines, the double point is called an *acnode* or *isolated point*. The Folium of Descartes, $x^3 + y^3 = 3axy$ (Fig. 7–6), has a node at the origin and the line $x + y + a = 0$ as an asymptote (Section 7–6), the semicubical parabola $y^2 = x^3$ (Fig. 7–7) has a cusp at the origin, and the curve $y^2 = x^2(x - 1)$ (Fig. 7–8) has an isolated point at the origin.

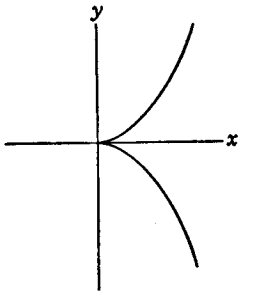


FIG. 7–7

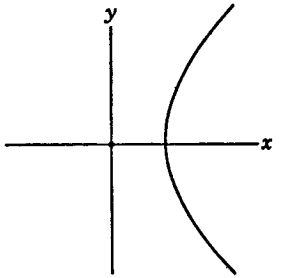


FIG. 7–8

Cubic curves are generally classified as follows, in terms of their double points:

(i) Cubic curves without double points: *elliptic cubics*,
(ii) Cubic curves with a node: *nodal cubics*,
(iii) Cubic curves with a cusp: *cuspidal cubics*.

Many of the properties of cubic curves may be found in [23; 139–243] and [26; 201–263].

Quartic curves may also be classified in terms of their singular points. Such a classification and many of the properties of quartic curves may be found in [23; 244–328] and [26; 264–349].

Plane curves of any degree n may be considered in terms of their double points and other singular points. In particular, any irreducible curve (Section 7–9) having its maximum number of singular points, i.e., *zero deficiency*, is called a *unicursal curve*. A unicursal curve is characterized by the fact that the coordinates of every point on the curve may be expressed rationally in terms of a single parameter. Unicursal curves are important in several mathematical theories.

The next two sections contain further details on the graphing of higher plane curves.

EXERCISES

1. Show that an irreducible cubic curve has at most one double point.
2. Graph an example of each of the following curves and give its equation: (a) elliptic cubic, (b) nodal cubic, (c) cuspidal cubic.

7–6 Rational functions. A polynomial $f(x, y)$ in x and y with complex coefficients may be considered as a polynomial in y with coefficients from the ring of polynomials in x with complex coefficients. The equation $f(x, y) = 0$ then defines y as an *algebraic function* of x (Section 3–16). If $f(x, y)$ is of degree n in y, there are exactly n complex values of y for each value of x such that the coefficient of y^n does not vanish. Thus an algebraic function is not, in general, single-valued for n greater than one. For $n = 1$ we have $f(x, y) = p(x)y - q(x) = 0$, where $p(x)$ and $q(x)$ are polynomials in x. In this section we shall consider the special case $y = q(x)/p(x)$, where $p(x)$ and $q(x)$ are relatively prime polynomials in x with real coefficients.

We shall be especially concerned with intercepts and asymptotes. The graph (Fig. 7–9) of the equation $2x + 3y = 6$ has x-intercept 3 and y-intercept 2. In general, each intersection of a curve with a coordinate axis may be taken with the origin and unit point to determine a segment of signed length equal to the coordinate of the point of intersection. This coordinate is called an *intercept* of the curve. In particular, the real zeros of a function $y = f(x)$ are the x-intercepts of its graph.

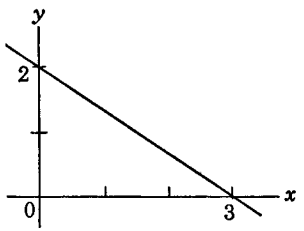


FIG. 7–9

An asymptote may also be easily described. Suppose a variable point P on the graph of $f(x, y)$ moves so that one or both of its coordinates becomes indefinitely large. If simultaneously the point P approaches indefinitely near to a line $ax + by + c = 0$, that line is called an *asymptote* of the graph of $f(x, y)$. For example, $x^2y - 1$ has both the coordinate axes as asymptotes (Fig. 7–10); $xy - 2y - 1$ has the lines $x = 2$ and $y = 0$ as asymptotes (Fig. 7–11). The asymptotes of the form $y = c$ are called *horizontal asymptotes;* those of the form $x = c$, *vertical asymptotes.* There also exist other asymptotes as, for example, the line $x + y + a = 0$ in Fig. 7–6.

The horizontal asymptotes of the graph of a polynomial $f(x, y)$ may be found by equating to zero the coefficients of the highest power of x in $f(x, y)$. Similarly, the vertical asymptotes may be found by equating to zero the coefficients of the highest power of y. For example, $xy - 2y - 1$ has horizontal asymptote $y = 0$ and vertical asymptote $x = 2$ (Fig. 7–11).

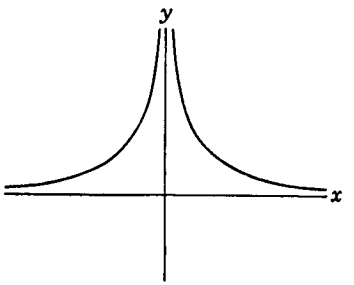


FIG. 7–10

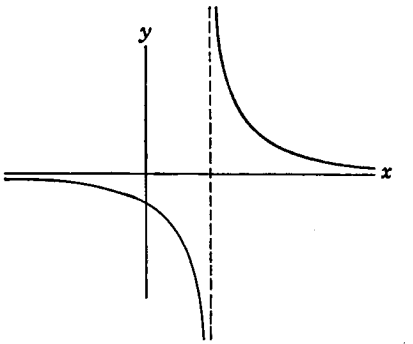


FIG. 7–11

The function $y = q(x)/p(x)$, where $q(x)$ and $p(x)$ are polynomials in x, is called a *rational function* of x (Section 3–3). We shall assume that $q(x)$ and $p(x)$ are relatively prime (Section 3–4). The graph of this rational function then has x-intercepts at real roots of $q(x)$ and vertical asymptotes corresponding to the real roots of $p(x)$. The horizontal asymptotes may also be readily found (Exercises 2 and 3). Let

$$q(x) = a_0x^n + a_1x^{n-1} + \cdots + a_n, \quad a_0 \neq 0,$$
$$p(x) = b_0x^m + b_1x^{m-1} + \cdots + b_m, \quad b_0 \neq 0.$$

If $n < m$, the graph of $q(x)/p(x)$ has the x-axis as an asymptote in both positive and negative senses. If $n = m$, the graph has the line $y = a_0/b_0$, as a horizontal asymptote in both senses. If $n > m$, there are no horizontal asymptotes. We may frequently also find other asymptotes of rational functions (Exercise 5). With these few rules and the fact that the function changes sign at a vertical asymptote $x = b$ if and only if the root $x = b$ is of odd multiplicity in $p(x)$, the graphing of the rational function $y = q(x)/p(x)$ is no more difficult than that of the polynomial $z = p(x) \cdot q(x)$. In fact, y and z have the same sign whenever $z \neq 0$, since $z = y[p(x)]^2$.

Consider the example

$$y = \frac{(x-2)(x^2-4)(2x-7)}{x(x-1)^2(x+3)}.$$

The curve has 2, 2, −2, and $\frac{7}{2}$ as x-intercepts, vertical asymptotes at $x = 0, 1, 1, -3$, horizontal asymptote $y = 2$. The general shape of the graph is given in Fig. 7–12. A more exact graph can be obtained by using the calculus to obtain the inflection points or simply by plotting a few additional points.

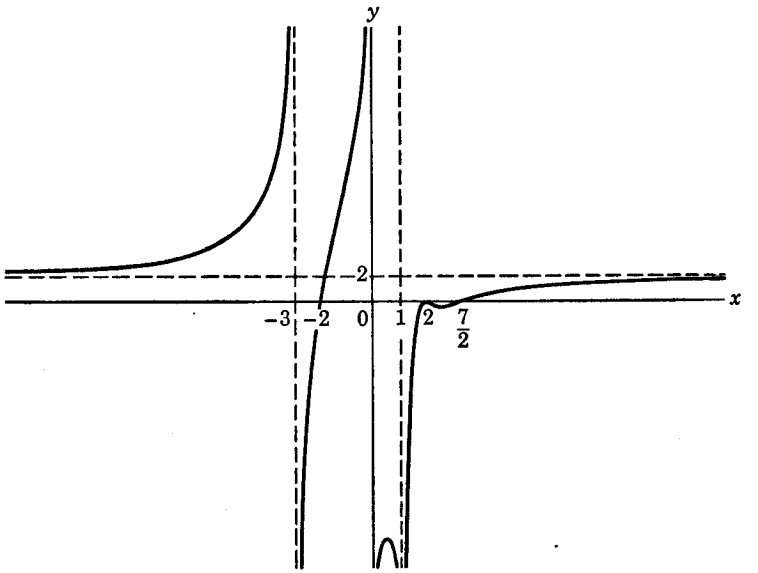


FIG. 7–12

This method may also be used to graph functions of the form $p(x)t^2 = q(x)$. First consider $p(x)y = q(x)$ and then let $t = \pm\sqrt{y}$. The graph in t and x will be real only for those values of x corresponding to positive or zero values of y. For example, in order to graph the function

$$f(x, t) = x(x-1)^2(x+3)t^2 - (x-2)(x^2-4)(2x-7) = 0,$$

we first graph the corresponding equation, where $y = t^2$, as in the example above. From the above graph it is evident that the graph of $f(x, t)$ is real only for values of x on the intervals $x < -3$, $-2 \leqq x < 0$, $x = 2$, and $7/2 \leqq x$. Then the point (2, 0) is an isolated point, the line $x = -3$ is a vertical asymptote, the lines $t = \pm\sqrt{2}$ are horizontal aymptotes, and the graph of $f(x, t)$ is of the form given in Fig. 7–13.

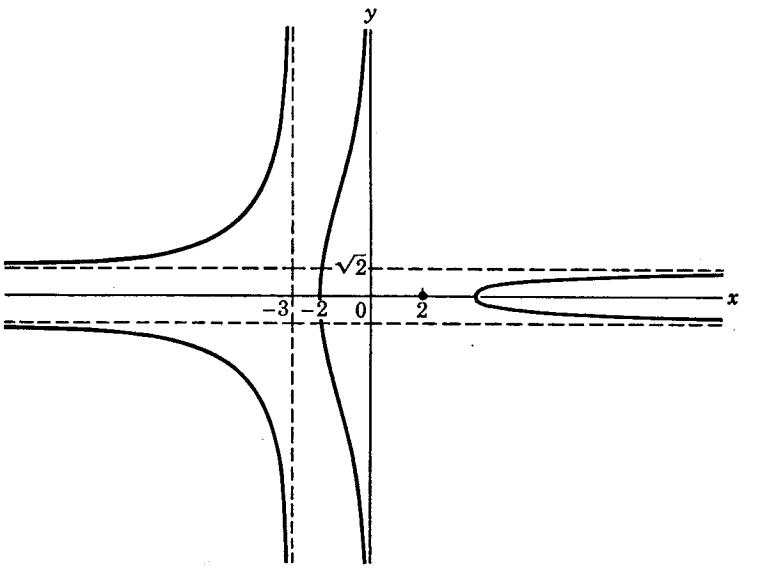


FIG. 7–13

EXERCISES

1. Give the x- and y-intercepts of the following curves:
   (a) $x/a + y/b = 1$,
   (b) $y = x^2 - 2x$,
   (c) $(x^2 - 4)y = x + 5$,
   (d) $x^2y - 2xy - x^2 + x + 6 = 0$,
   (e) $(x^2 - 2x - 3)y = x^3 - 2x^2 + x$,
   (f) $(x + 1)y^2 = x(x - 2)^2(x - 3)$.

2. Consider $y = q(x)/p(x)$ in the form $yp(x) = q(x)$ and discuss the horizontal and vertical asymptotes, using the coefficients of the highest powers of the variables.

3. The horizontal asymptote of $y = q(x)/p(x)$ may be defined as $y = \lim_{x\to\infty} q(x)/p(x)$ whenever the limit (finite) exists. Use the methods of Section 3–11 to verify that the statements made regarding horizontal asymptotes in Exercise 2 are also valid under this new definition.

4. Find the vertical and horizontal asymptotes of the curves in Exercise 1, whenever such exist.

5. Write $y = q(x)/p(x)$ in the form $s(x)/p(x) + r(x)$ where the degree of $s(x)$ is less than that of $p(x)$ and prove that the graph of the given rational function is asymptotic to the graph of the polynomial $y = r(x)$.

7-7 Algebraic functions. A polynomial equation $f(x, y) = 0$, where $f(x, y)$ is considered as a polynomial in y with coefficients from the ring of polynomials in x with complex coefficients, defines y as an algebraic function (Section 3–16) and has as its real graph an *algebraic plane curve.* The concept of an algebraic function is an extension of the concept of a polynomial in that every polynomial $p(x)$ satisfies $f(x, y) = y - p(x)$, that is, a polynomial is a special case of an algebraic function that arises when $f(x, y)$ has the form $y - p(x)$. Similarly, a rational function (Section 7–6) is a special case of an algebraic function that arises when $f(x, y)$ has the form $p(x)y - q(x)$.

We have seen (Section 3–10) that any polynomial $p(x)$ is a single-valued function of x. Also, a rational function of x is single-valued whenever it is defined (Section 3–3). However, an algebraic function defined by a polynomial equation $f(x, y) = 0$ of degree n in y may have n values of y corresponding to a given value of x. For example, $y^2 - x = 0$ associates two real values of y with each positive value of x. Our discussion in this section consists of a very brief introduction of a graphical representation (the Riemann surface) of the n values of the algebraic function y (not necessarily real or distinct) corresponding to each value of x (Theorem 4–2).

We have often spoken of a polynomial with real coefficients as a "real polynomial." In some texts a *curve* is said to be real if the function $f(x, y)$ has real coefficients. With this definition a real curve may have only imaginary points, as in the case of $x^2 + y^2 + 1$. We shall adopt the terminology that the *real graph* of any curve is composed of the real points whose coordinates (real n-tuples) make the function vanish.

When considering our number system in Chapter 1, we extended the rational number system to the real number system in order to obtain a continuous system (no gaps in the axis of real numbers). Then we extended the real number system to the complex number system with the property that every polynomial of degree n in a single variable with complex coefficients has exactly n complex roots (Theorem 4–2), i.e., the complex number system is algebraically closed.

Analogous geometric statements can be made as follows. We considered a line of rational points in order to have the roots of all linear equations with integral or rational coefficients. This line was extended to the real line for the sake of continuity, i.e., so that any

curve joining points on "opposite sides" of a line will intersect the line. Finally, the real line was extended to the complex plane in order to represent all the roots of any polynomial in one variable with complex coefficients.

A similar extension may be made geometrically in considering the graphs of algebraic functions. In a space with two complex coordinates, the equation $y^2 - x = 0$ associates with every complex value of x exactly two values of y. These values of y are real and distinct when x is positive, real and equal when x is zero, conjugate imaginary when x is negative. Visualization of such a space is difficult, since it corresponds to a four-dimensional Euclidean space. The complex variable x is essentially defined over a Euclidean plane. The two values of y may be identified with two surfaces or sheets $y = +\sqrt{x}$ and $y = -\sqrt{x}$. On this new surface of two sheets, called the *Riemann surface* of the function $y^2 - x = 0$, the algebraic function $y = f(x)$ is single-valued. The sheets of a Riemann surface often cross themselves at points (called branch points), since two planes in four-space may intersect in a single point. Even though this property makes visualization difficult, the concept of sheets has been found to be helpful.

In general, there is associated with every algebraic function $y(x)$, defined by a polynomial equation $f(x, y) = 0$ of degree n in y, a Riemann surface of n sheets on which the function is single-valued. Essentially, the Riemann surface is the graph of the function in a space with two complex coordinates. Points on the Riemann surface in the neighborhood of the points corresponding to $x = a$ may be studied using infinite series expansions (Section 3–11) in a variable t, where $x = a + t^r$ [8; 32]. A thorough treatment of this rather intricate procedure may be found in [8].

The remainder of this chapter is devoted to a few practical procedures in the construction of real graphs, graphical solution of algebraic equations, and the determination of empirical equations.

EXERCISES

1. Give three of each of the following types of algebraic functions of x: (a) polynomial, (b) rational function but not a polynomial, (c) algebraic function but not a rational function.

2. Indicate the number of sheets in the Riemann surface of each of the functions written down in Exercise 1.

7–8 Curve tracing. The tracing of the curve for a given equation may be performed by constructing the curve mechanically, by plotting numerous points with a certain degree of accuracy, or by plotting a few selected points and determining the shape of the curve from its symmetry, singular points, and asymptotes. We shall consider each of these methods briefly.

Theoretically, the problem of mechanically constructing the real graph of any algebraic plane curve $f(x, y) = 0$ of degree n was completely solved by Alfred Bray Kempe using linkages (Section 6–9). In practice, the linkages often become very complicated and cumbersome as n increases. However, many common plane curves may be conveniently constructed using linkages. Yates [54] indicates how to construct many curves mechanically, including the cardioid, cassinian, cissoid, conic sections, Lemniscate of Bernoulli, and Limaçon of Pascal. We all use compasses to draw a circle and, possibly, a loop of string around pegs at the foci to draw an ellipse. Some people find a pantograph (Exercise 4, Section 6–9) and a few other mechanical devices very useful. However, most mechanical methods require too much equipment and often very special devices for ordinary use.

The plotting of a curve of degree n by means of a large number of points can be difficult as well as tedious. The difficulties arise algebraically in that the roots of polynomial equations of degree n must be approximated, and they arise geometrically in that the sequence of points on the curve or the association of the points plotted may not be immediately obvious. In general, the method of simply plotting points is most useful for unicursal curves (Section 7–5).

Given the equation $f(x, y) = 0$ of the curve to be graphed, there are several methods for simplifying the procedure of plotting numerous points. In the first place, the points at which the curve intersects the coordinate axes are easy to plot and are usually at least as easy to obtain as any others: the x-intercepts are the real zeros of $f(x, 0)$ and the y-intercepts are the real zeros of $f(0, y)$. For example, $x^2 + y^2 - 5x + y - 6$ has $f(x, 0) = x^2 - 5x - 6$, whence the x-intercepts are 6 and -1; $f(0, y) = y^2 + y - 6$, whence the y-intercepts are 2 and -3.

Secondly, the curve may be symmetrical to certain axes or points, so that only a portion of the graph need be carefully plotted and the remainder can be obtained by symmetry. For example, the graph

of $y - x^2$ is symmetrical to the y-axis. In general, the graph of $f(x, y)$ is symmetrical

to the x-axis if $f(x, y) = f(x, -y)$,
to the y-axis if $f(x, y) = f(-x, y)$,
to the origin if $f(x, y) = f(-x, -y)$,
and to the line $y = x$ if $f(x, y) = f(y, x)$.

Tests for symmetry with respect to many other axes and centers may be developed (Exercise 3), but the above are the most common and the easiest to apply (Exercise 2).

Another concept that simplifies the plotting of points is that of excluded regions. Frequently there exist sets of values of a variable for which the graph has no real value. Such sets of values are called *excluded regions*. For example, the graph of $y = x^2$ has no real points for negative values of y; the graph of $y^2 = 2x - x^2$ has no real points when $x < 0$ or $2 < x$; the graph in Fig. 7–13 has no real points when $-3 \leqq x < -2$, $0 \leqq x < 2$, or $2 < x < \frac{7}{2}$.

Frequently the behavior of a graph near the origin or some other point is of special interest. By means of a translation $x' = x - a$, $y' = y - b$ the behavior at any point (a, b) may be studied as the behavior at the new origin. In general, the terms of least degree determine the behavior of the curve near the origin. Given any polynomial $f(x, y)$, let $g(x, y) = 0$ be the polynomial equation obtained by equating the terms of lowest degree in $f(x, y)$ to zero. If $g(x, y)$ is constant and different from zero, the graph of $f(x, y)$ does not pass through the origin. In any case all terms of $g(x, y)$ are of the same degree, that is, $g(x, y)$ is a *homogeneous polynomial*. A homogeneous polynomial of degree r in x and y with complex coefficients may always theoretically be expressed as the product of r linear polynomials with complex coefficients. The graphs of the linear factors of $g(x, y)$ are the tangents to the curve $f(x, y) = 0$ at the origin. For example, if $f(x, y) = x^3 + y^3 - 3axy$, where $a \neq 0$ (Fig. 7–6), then $g(x, y) = -3axy$ and the tangents at the origin are $x = 0$ and $y = 0$. Similarly, the graph of $x^3 + xy^2 + ax^2 - ay^2 = 0$ has tangents $x + y = 0$ and $x - y = 0$ at the origin.

We found horizontal and vertical asymptotes very useful in graphing rational functions (Section 7–6). In general, given any polynomial $f(x, y)$ of degree n, the terms of degree n and $n - 1$ determine the behavior of the curve for large numerical values of the coordinates. Let $h(x, y)$ be the homogeneous polynomial consisting of the

terms of $f(x, y)$ of degree n. The real graphs of the linear factors of $h(x, y)$ are parallel to the asymptotes of the graph of $f(x, y)$. In the example $f(x, y) = x^3 + y^3 - 3axy$ (Fig. 7–6), $h(x, y) = x^3 + y^3$, and the only asymptote is parallel to $x + y = 0$. In general, the equations of the asymptotes depend upon the terms of degrees n and $n - 1$ [27; 13]. If $h(x, y)$ does not contain a term x^n, let $p(y)$ be the coefficient of the highest power of x occurring in the original polynomial $f(x, y)$. Then the graphs of the linear factors of $p(y)$ are the horizontal asymptotes of $f(x, y)$. A similar statement can be made regarding the vertical asymptotes, as illustrated by the rational functions (Section 7–6) when considered in the form $p(x)y - q(x)$. The general theory of asymptotes involves the Newton Diagram or the analytical triangle [27; 15].

We have now discussed the use of intercepts, symmetry, excluded regions, tangents at the origin (or any other specified point), and asymptotes in tracing the graph of a polynomial $f(x, y)$. The singular points (Section 7–5) may also be effectively used. Frost [22] and Johnson [27] treat this subject in an elementary manner and do not assume any calculus. Hilton [26] makes very effective use of more advanced mathematical concepts. We shall conclude this section with a single example of the powerful methods used in more advanced texts, such as [26].

Given any polynomial $f(x, y)$ of degree n, we may make each term of the polynomial of degree n by inserting a suitable power of z and thereby obtain a homogeneous polynomial $f(x, y, z)$. For example, if $f(x, y) = x^3 + y^3 - 3xy$, then $f(x, y, z) = x^3 + y^3 - 3xyz$. We now use the notation f_x to indicate the first partial derivative with respect to x of $f(x, y, z)$, that is, the first derivative of $f(x, y, z)$ with respect to x where y and z are considered as constants. Similarly, f_{xy} is the partial derivative of f_x with respect to y, etc. In the case of $f(x, y, z) = x^3 + y^3 - 3xyz$, we have

$$\begin{array}{lll} f_x = 3x^2 - 3yz, & f_y = 3y^2 - 3xz, & f_z = -3xy, \\ f_{xx} = 6x, & f_{yy} = 6y, & f_{zz} = 0, \\ f_{xy} = f_{yx} = -3z, & f_{yz} = f_{zy} = -3x, & f_{zx} = f_{xz} = -3y. \end{array}$$

The determinant of the second order partial derivatives of $f(x, y, z)$,

$$H(x, y, z) = \begin{vmatrix} f_{xx} & f_{xy} & f_{xz} \\ f_{yx} & f_{yy} & f_{yz} \\ f_{zx} & f_{zy} & f_{zz} \end{vmatrix},$$

has as its graph the *Hessian* [26; 98] of the given curve. The importance of the Hessian arises from the fact that the intersections of a curve with its Hessian are precisely the singular points and inflection points of the given curve. For example, the Hessian of $x^3 + y^3 - 3xyz$ is the graph of the polynomial

$$H(x, y, z) = \begin{vmatrix} 6x & -3z & -3y \\ -3z & 6y & -3x \\ -3y & -3x & 0 \end{vmatrix} = -54(x^3 + y^3).$$

The intersections of the graphs of $f(x, y) = x^3 + y^3 - 3xy$ and $H(x, y) = -54(x^3 + y^3)$ must also lie on $54f(x, y) + H(x, y)$ and therefore on at least one of the coordinate axes. Thus the only singular point or inflection point of the graph of $f(x, y) = x^3 + y^3 - 3xy$ is at the origin (Fig. 7–6).

In this section we have considered several methods for graphing plane curves. However, we have in no sense exhausted the subject, since one can easily write down the equation of a higher plane curve that cannot be readily graphed by these methods. Thus, recognizing the difficulty of graphing most higher plane curves, we shall consider the above discussion as a brief treatment of methods that may facilitate the graphing of any given higher plane curve but cannot be expected to make graphing a simple routine. In the next section we shall consider a few special types of graphs.

Exercises

1. Find the x-intercepts and the y-intercepts of each of the following curves:

(a) $x^2 - 8x + y^2 + 7y + 12 = 0$,
(b) $x^3 + y^3 - 3axy = 0$ (Fig. 7–6), Folium of Descartes,
(c) $x^2y^2 + x^2 - y^2 = 0$,
(d) $xy^2 - x - 4y = 0$,
(e) $x^2y^2 - x^2 - 4y^2 = 0$,
(f) $x^2y + a^2y - a^3 = 0$, Witch of Agnesi,
(g) $x^3 + xy^2 - 2ay^2 = 0$, Cissoid of Diocles,
(h) $x^3 + xy^2 + ay^2 - 3ax^2 = 0$ (Fig. 6–5), Trisectrix,
(i) $x^3 + xy^2 + ax^2 - ay^2 = 0$, Strophoid,
(j) $x^2y + b^2y - a^2x = 0$, Serpentine,
(k) $y^2 = x(x - 1)(x - 2)$,
(l) $y^2 = -x^2(x + 1)(x - 2)$.

2. Test each of the curves in Exercise 1 for symmetry with respect to the (a) x-axis, (b) y-axis, (c) origin, (d) line $x = y$.

3. Develop tests for symmetry with respect to each of the following lines: (a) $x = 1$, (b) $x = a$, (c) $y =$ b, (d) $x - y = 0$.

4. Indicate the excluded regions, if any exist, for x and y in the graphs of each of the curves in Exercise 1.

5. Find the tangents to the following curves at the origin:

(a) $xy^2 = x^3$, (c) $(x + 2)y^2 = x^2(x + 1)$,
(b) $y^2 = x^2(x - 1)$, (d) $xy^2 - x^3 = x^4$.

6. Find the tangents at the origin, if any exist, for the graphs of each of the curves in Exercise 1.

7. Discuss the behavior of $x^3 + y^3 - 6xy$ at (3, 3).

8. Use the Hessian to find the singular points of each of the following curves: (a) $x^3 - xy^2 - y^2$, (b) $x^3 + xy^2 - x$.

9. Graph the curves in Exercise 8. [*Hint:* The curve in Exercise 8(b) is reducible (Section 7–9).]

7–9 Special graphs. We now consider four special methods that may be used in graphing. We shall consider graphs of reducible curves, graphs obtained by the composition of ordinates, graphs of transcendental functions, and Peano's space-filling arc.

Whenever the polynomial $f(x, y)$ may be factored as

$$f(x, y) = g(x, y) \cdot h(x, y),$$

the graph of $f(x, y)$ is said to be *reducible* and consists of the totality of points on the graphs of $g(x, y)$ and $h(x, y)$. The graphs of the factors of $f(x, y)$ are called *components* of the graph of $f(x, y)$. For example, the graph of $x^3 + xy^2 - x$ [Exercise 8(b), Section 7–8] has two components corresponding to the factors x and $x^2 + y^2 - 1$, respectively. Thus the graph of any reducible curve may be obtained by graphing each of its components.

The graphing of $f(x, y)$ may also often be simplified by solving for one of the variables, say $y = r(x)$, where $r(x)$ is not necessarily a polynomial. For example,

$$2x^2 + y^2 - 2xy + 2x - 2,$$

when solved for y, gives

$$y = x \pm \sqrt{2 - 2x - x^2},$$

which can be graphed as $y = y_1 + y_2$, where $y_1 = x$ and

$$y_2 = \pm \sqrt{2 - 2x - x^2}$$

or $(x + 1)^2 + y_2^2 = 3$. When $f(x, y) = 0$ is solved for y, this method of plotting the curve is referred to as graphing by the *composition of ordinates.*

Throughout our discussion, we have been primarily concerned with graphs of polynomials. Graphs may also be defined by transcendental functions such as $y = \sin x$ or $y = \log_p x$, where $1 < p$, by several polynomials

$$\begin{aligned} y &= -x \text{ when } x < 0, \\ &= 5 \quad \text{when } x = 0, \\ &= x^2 \quad \text{when } x > 0, \end{aligned}$$

or by other symbols such as $y = |x|$ or $y = [x]$, where the bracketed x indicates the greatest integer less than or equal to x. Several such functions are discussed in [24; 55–60]. It is even possible to define y so that its graph is very misleading, due to the breadth of the marks corresponding to the points plotted. For example, suppose

$$\begin{aligned} y &= 1 \text{ when } x \text{ is rational}, \\ &= 0 \text{ when } x \text{ is irrational}. \end{aligned}$$

Since the rational numbers are dense and the irrational numbers are uncountable (Section 1–13) on every segment (a, b), $a < b$, the graph of the above single-valued function appears to be the two lines $y = 0$ and $y = 1$. Several other examples of graphs that cannot be given by a single polynomial in x and y may be found in Exercise 3.

We now conclude our discussion of the tracing of graphs by mentioning an arc with a very unusual property. This arc, Peano's space-filling arc, passes through every interior point of a square. It is called an arc because each point on it may be specified by a single real value of the parameter t, $0 < t < 1$, just as the interior points of a unit segment on the x-axis may be specified by a single real value of x, $0 < x < 1$. A description of the manner in which the interior points of the square are associated with real values of t, $0 < t < 1$, may be found in [21; 56]. The arc has importance in mathematical theories, since it gives a one-to-one correspondence between the points of an element of area (two-dimensional) and the points of a line segment (one-dimensional).

Exercises

1. Graph the following:
   (a) $x(x^2 + y^2 - 2x + 2y - 7)$,
   (b) $(x^2 - y^2)(x^2 + y^2 - 1)$,
   (c) $x^3 - x^2y - 2x + 2y$.

2. Graph the following by composition of ordinates or abscissas:
   (a) $y = x^3 + 1$,
   (b) $x^2 - 4xy + 4y^2 = x - 4$,
   (c) $x^2 + 6xy + 9y^2 + x + 6y + 1 = 0$,
   (d) $y = x + \sin x$.
3. Graph the following:
   (a) $y = x - [x]$.
   (b) $y = |x|$.
   (c) $y = -x$ when $x \leqq 0$,
   $= x$ when $x > 0$.
   (d) $y = \sqrt{x}$ when x is the square of an integer,
   $= x$ otherwise.
   (e) $y = 1$ when x is rational,
   $= -1$ when x is irrational.
   (f) $y^2 = |x|$.
   (g) $y^2 = x - [x]$.

7–10 Graphical solutions. If the solution of a problem is obtained by observing the intersections of lines and curves on a plane, it is necessarily an approximate solution. However, an approximate solution is frequently very useful. The following graphical addition is a trivial example of the general procedure.

Consider the family F of lines $x + y = C$ and make a chart by graphing the lines corresponding to values of C less than some positive number N in absolute value. The sum $a + b$ of two real numbers may now be found graphically by plotting the point (a, b) and observing the particular value of C that would correspond to the line through (a, b) of the family of lines F.

The applications of graphical methods are very numerous. Addition, subtraction, multiplication, and division, compound and simple interest, solution of quadratic, cubic, and quartic equations, integration and differentiation are considered in [46]. The use of curves in the trisection problem has been discussed in Section 6–7. Many other applications of graphical methods may be found in [33].

Another application of graphical solutions is found in the use of slide rules and nomograms. For example, on most slide rules one may find the square root of a number on the A scale by looking directly below it (assuming the ends of the scales correspond) on the C scale. The correspondence between the number and its square root is usually observed by means of a cross-hair (movable narrow line perpendicular to the scales). The slide rule is thus a special case of an alignment chart or nomogram. Many nomograms consist of three precisely located lines or curves with scales (possibly very different in units and meaning) marked upon each. The nomogram

is then used by placing a straightedge through points on two of the scales according to given data, and reading the result from the third scale. Maurice d'Ocagne [42] devised some excellent nomograms and developed methods for handling equations in more than two variables on the plane. Many of the procedures in more recent books are based upon the work of d'Ocagne. The reader is referred to special texts on the subject for details regarding the construction and use of nomograms. We shall conclude our discussion of graphical representations with a brief discussion of the problem of finding an equation or curve to fit a given set of points.

Exercises

1. Draw graphs that may be used to approximate each of the following when $0 \leqq b \leqq 10$:
   (a) $\sqrt{b}$,
   (b) $b^2 + 1$,
   (c) $\sqrt[4]{b}$,
   (d) $b^{\frac{2}{3}}$,
   (e) $\sqrt{b + 3}$,
   (f) $\sqrt{b} + 3$.
2. Draw several curves in a family of curves that may be used to approximate each of the following:
   (a) $a + b$,
   (b) $a - b$,
   (c) ab,
   (d) a/b,
   (e) $a^2 + 2ab + b^2$.
3. Solve the following systems of equations and inequalities graphically:
   (a) $y = x$, $y < x^2$, $y^2 > x$.
   (b) $1 - y^2 > x^2$, $0 < x < y$.
   (c) $\sin x < y \leqq 1$, $|x| < 3$.
   (d) $1/x > 1/y$, $x^2 + y^2 = 25$.

7–11 Curve fitting. We have previously constructed graphs for given equations. In this section we assume that a set of corresponding values of two variables is given and endeavor to find the equation between the variables that best fits the given data and the conditions of the problem. Often the data is obtained by observation or experiment, and new insight into the problem can be gained from an equation between the variables.

It is always possible to find a curve of degree less than or equal to $n - 1$ that passes through n given points. However, the graph may be such that additional sets of values from the graph do not approximate the corresponding new pairs of values obtained from the problem. We shall mention a few methods for determining an

equation, called an *empirical equation*, that the given data approximately satisfies. Since a straight line is one of the easiest curves to visualize, we shall consider it on several types of coordinate scales. The suitability of an empirical equation is often measured by means of averages or by the method of least squares, i.e., by minimizing the sum of the squares of the differences between the results in the given data and those arising from the equation. The topic of curve fitting is considered in many analytic geometry texts, for example, [38; 204–223]. More extensive treatments are found in [14; 3–88] and in [33; 120–169].

The graph of $y = ax + b$ is a straight line with slope a and y-intercept b on ordinary graph paper. Conversely, whenever the points corresponding to the given data appear to lie on a straight line on ordinary graph paper, they can (by a change of coordinates if the line is parallel to the y-axis) be closely fitted by an equation of the form $y = ax + b$. The constants a and b can be readily calculated from the given data or from the graph of the line.

The method of averages may be applied to the linear relation $y = ax + b$ as follows. The given pairs of values, for example,

x	1	2	3	4	5	6	7
y	3.20	4.30	4.80	6.10	7.00	8.20	9.10

are divided into two sets (since two constants are to be determined) approximately equal in number. For example, take the first three and last four pairs in the above data. The pairs of each set are substituted into the linear relation, giving the *rectifying equations*

$$\begin{aligned} 3.2 &= a + b, & 6.1 &= 4a + b, \\ 4.3 &= 2a + b, & 7.0 &= 5a + b, \\ 4.8 &= 3a + b, & 8.2 &= 6a + b, \\ & & 9.1 &= 7a + b; \end{aligned}$$

the elements of each set are added,

$$12.3 = 6a + 3b, \quad 30.4 = 22a + 4b,$$

and the resulting two equations are solved simultaneously for $a = 1$ and $b = 2.1$ to give the empirical equation $y = x + 2.1$. The method of averages may also be used to fit a parabola $y = ax^2 + bx + c$ to the data by dividing the pairs of values into three sets and solving the three resulting linear equations for the three parameters a, b, c.

The method of least squares, when applied to the linear relation $y = ax + b$, essentially gives values of a and b such that

$$D = \sum_{j=1}^{n} (ax_j + b - y_j)^2$$

is a minimum where there are n pairs of values (x_j, y_j), $j = 1, 2, \ldots, n$, in the given data. This may be done by solving the two linear differential equations $\delta D/\delta a = 0$, $\delta D/\delta b = 0$ simultaneously for a and b. It may also be done [38; 219–222] by solving simultaneously the equations

$$\sum y_j = a \sum x_j + nb,$$
$$\sum x_j y_j = a \sum x_j^2 + b \sum x_j.$$

If the rectifying equations $y_j = ax_j + b$ form one column and the $x_j y_j = ax_j^2 + bx_j$ another, we have for the above example

$$\begin{aligned} 3.2 &= a + b & 3.2 &= a + b \\ 4.3 &= 2a + b & 8.6 &= 4a + 2b \\ 4.8 &= 3a + b & 14.4 &= 9a + 3b \\ 6.1 &= 4a + b & 24.4 &= 16a + 4b \\ 7. &= 5a + b & 35. &= 25a + 5b \\ 8.2 &= 6a + b & 49.2 &= 36a + 6b \\ 9.1 &= 7a + b & 63.7 &= 49a + 7b \\ \hline 42.7 &= 28a + 7b & 198.5 &= 140a + 28b \end{aligned}$$

whence $a = 0.99$ and $b = 2.14$, to give the empirical equation $y = 0.99x + 2.14$. These methods clearly apply, with slight modifications, to the linear relations $\log y = a \log x + \log b$ and $\log y = ax + b$, which we now consider.

The equation $y = bx^a$ is equivalent to $\log y = a \log x + \log b$ that is linear in $\log y$ and $\log x$. Thus logarithmic paper is used, on which the distances Ox and Oy represent the logarithms of x and y respectively instead of their magnitudes. Whenever the points corresponding to the given data appear to lie on a straight line on logarithmic paper, they can be closely fitted by an equation of the form $y = bx^a$ and the constants may be readily determined from the graph. This method may be extended to include $y = bx^a + c$ by considering it in the form $\log (y - c) = a \log x + \log b$. The value of c may be calculated by trial and error or graphically [14; 12].

We observe also that $y = 10^{b+ax} + c$ may be considered in the form $\log (y - c) = ax + b$ as a linear graph on semilogarithmic paper. Finally, when the rate of change of one variable with respect to the other is linear, we have

$$\frac{\Delta y}{\Delta x} = 2ax + b$$

and the variables may be related by

$$\frac{dy}{dx} = 2ax + b \quad \text{or} \quad y = ax^2 + bx + c.$$

If the reciprocals of x and y are linearly related, i.e.,

$$\frac{1}{y} = \frac{b}{x} + a,$$

then

$$y = \frac{x}{ax + b}.$$

There are other common types of empirical equations, but the above indicate one of the uses of various coordinate papers and some of the advantages of graphical methods. More extensive treatments of empirical equations may be found in [14] and many texts on statistics.

EXERCISES

1. Use the method of averages to find an equation of the form $y = ax + b$ that approximately fits the following data:

x	1	2	3	5	10	12
y	2.2	4.3	6.5	11	21	24

2. Plot the given points and the line obtained in Exercise 1.

3. Find an equation of the form $y = ax^2 + bx + c$ for the data given in Exercise 1.

7–12 Conclusion. Throughout this chapter we have considered the relationships between graphs and equations. In Sections 7–1 to 7–10 our primary concern has been to obtain the graph of a given function; in Section 7–11, to find an equation of a certain type having a curve that best fits the points corresponding to certain given data. These considerations illustrate the importance of the relationship between the fundamental concepts of algebra and the fundamental concepts of geometry. In particular, we have seen that many algebraic problems may be expressed geometrically and, similarly, many geometric problems (for example, the classical constructions in Chapter 6) may be expressed algebraically. This illustrates the fact that a study of the basic concepts of any branch of mathematics, and in particular our study throughout this text of the fundamental concepts of algebra, increases one's understanding of all phases of mathematics.

BIBLIOGRAPHY

1. ALBERT, A. A., *College Algebra.* New York: McGraw-Hill, 1946.

2. ARCHIBALD, R. C., "Outline of the History of Mathematics," 6th ed., *American Mathematical Monthly*, **56**, Part II, 1949.

3. BALL, W. W. R., *Mathematical Recreations and Essays.* Eleventh ed., revised by H. S. M. Coxeter. New York: Macmillan, 1939.

4. BERGER, E. J., "Devices for a Mathematical Laboratory," *The Mathematics Teacher*, **44**, 34, 1951.

5. BERGER, E. J., "Devices for a Mathematical Laboratory," *The Mathematics Teacher*, **45**, 287, 1952.

6. BIRKHOFF, G. D. and BEATLEY, R., *Basic Geometry.* Chicago: Scott, Foresman and Co., 1940.

7. BIRKHOFF, G. and MACLANE, S., *A Survey of Modern Algebra.* New York: Macmillan, 1941.

8. BLISS, G. A., *Algebraic Functions.* Colloquium Publications, Vol. 16. New York: American Mathematical Society, 1933.

9. BÔCHER, MAXIME, *Introduction to Higher Algebra.* New York: Macmillan, 1907.

10. BOURBAKI, NICHOLAS, "The Architecture of Mathematics," *American Mathematical Monthly*, **57**, 221–232, 1950.

11. COOLIDGE, J. L., *A History of the Conic Sections and Quadric Surfaces.* Oxford: Clarendon Press, 1945.

12. COURANT, R., *Differential and Integral Calculus.* Vol. 1. E. J. McShane, translator. New York: Nordemann Publishing Co., Inc., 1938.

13. COURANT, R. and ROBBINS, H., *What is Mathematics?* New York: Oxford University Press, 1941.

14. DAVIS, D. S., *Empirical Equations and Nomography.* New York: McGraw-Hill, 1943.

15. DICKSON, L. E., "Constructions with Ruler and Compasses," *Monographs on Topics of Modern Mathematics.* J. W. A. Young, Editor. New York: Longmans, Green and Co., 1911, pp. 351–386.

16. DRESDEN, ARNOLD, *Solid Analytical Geometry and Determinants.* New York: Wiley, 1930.

17. DUBISCH, ROY, *The Nature of Number.* New York: Ronald Press, 1952.

18. EISENHART, L. P., *Coordinate Geometry.* Boston: Ginn, 1939.

19. FINE, H. B., *College Algebra.* Boston: Ginn, 1904.

20. FOURREY, E., *Procédés Originaux de Constructions Géométriques.* Paris: Libraire Vuibert, 1924.

21. FRANKLIN, PHILIP, *A Treatise on Advanced Calculus.* New York: Wiley, 1940.

22. FROST, PERCIVAL, *An Elementary Treatise on Curve Tracing.* 2nd ed. London: Macmillan, 1911.

23. GANGULI, SURENDRAMOHAN, *Lectures on the Theory of Plane Curves.* Parts I and II. Calcutta: University of Calcutta, 1919.

24. HARDY, G. H., *A Course of Pure Mathematics.* 9th ed. Cambridge: University Press, 1945.

25. HILSENRATH, JOSEPH, "Linkages," *The Mathematics Teacher,* **30,** 277–284, 1937.

26. HILTON, HAROLD, *Plane Algebraic Curves.* 2nd ed. London: Oxford University Press, 1932.

27. JOHNSON, W. W., *Curve Tracing in Cartesian Coordinates.* New York, Wiley, 1884.

28. KAMKE, E., *Theory of Sets.* F. Bagemihl, translator. New York: Dover, 1950.

29. KASNER, E. and NEWMAN, J., *Mathematics and the Imagination.* New York: Simon and Schuster, 1940.

30. KLEIN, FELIX, *Famous Problems of Elementary Geometry.* W. W. Beman and D. E. Smith, translators. Boston: Ginn, 1897.

31. LANDAU, EDMUND, *Foundations of Analysis.* F. Steinhardt, translator. New York: Chelsea, 1951.

32. LIEBER, L. R. and LIEBER, H. G., *Galois and the Theory of Groups.* Lancaster, Pa.: Science Press, 1932.

33. LIPKA, JOSEPH, *Graphical and Mechanical Computations.* New York: Wiley, 1918.

34. MCCOY, NEAL H., *Rings and Ideals.* Carus Mathematical Monograph, No. 8. Buffalo, N.Y.: Mathematical Association of America, 1948.

35. MESERVE, B. E., *Fundamental Concepts of Geometry.* Cambridge, Mass.: Addison-Wesley, 1953.

36. MESERVE, B. E., "Linkages as Visual Aids," *The Mathematics Teacher,* **39,** 372–379, 1946.

37. MESERVE, B. E., "The Euclidean Division Algorithm," *Pi Mu Epsilon Journal,* **1,** 138–144.

38. MIDDLEMISS, R. R., *Analytic Geometry.* New York: McGraw-Hill, 1945.

39. MUIR, THOMAS, *The Theory of Determinants in the Historical Order of Development.* 4 vols. London: Macmillan, 1906, 1911, 1920, 1923.

40. NAGELL, TRYGVE, *Introduction to Number Theory.* New York: Wiley, 1951.

41. NATIONAL COUNCIL OF TEACHERS OF MATHEMATICS, *Multi-Sensory Aids in the Teaching of Mathematics.* 18th yearbook. New York: Columbia University, 1945.

42. OCAGNE, MAURICE DE, *Traité de Nomographie.* Paris: Gauthier-Villars, 1899.

43. ORE, OYSTEIN, *Number Theory and Its History.* New York: McGraw-Hill, 1948.

44. PERLIS, S., *Theory of Matrices.* Cambridge, Mass.: Addison-Wesley Press, 1952.

45. ROOS, J. D. C. DE, *Linkages: the Different Forms and Uses of Articulated Links.* New York: D. Van Nostrand, 1879.

46. RUNNING, T. R., *Graphical Mathematics.* New York: Wiley, 1927.

47. THOMAS, J. M., *Theory of Equations.* New York: McGraw-Hill, 1938.

48. THOMAS, J. M., "Sturm's Theorem for Multiple Roots," *National Mathematics Magazine,* **15,** 391–394, 1941.

49. USPENSKY, J. V., *Theory of Equations.* New York: McGraw-Hill, 1948.

50. USPENSKY, J. V. and HEASLET, M. A., *Elementary Number Theory.* New York: McGraw-Hill, 1939.

51. VANDIVER, H. S., "Fermat's Last Theorem, Its History and the Nature of Known Results Concerning It," *American Mathematical Monthly,* **53,** 555–578, 1946.

52. VEBLEN, OSWALD and YOUNG, J. W., *Projective Geometry.* Vol. 2. Boston: Ginn, 1918.

53. WAERDEN, B. L. VAN DER, *Modern Algebra.* Vol. 1. Fred Blum, translator. New York: Ungar, 1949.

54. YATES, R. C., *Curves.* New York: Department of Mathematics, United States Military Academy, 1946.

55. YATES, R. C., *The Trisection Problem.* Baton Rouge: Franklin Press, 1942.

Symbols and Notation

The symbols and notation listed first appear and are defined on the pages noted at the left.

Page		Page	
3,6	$=$	88	$\phi(m)$
3,14	$<$	99	$p(x)$
3,14	$>$	114	$f(x)$
8	$\neq$	116	$[x]$
9	a^+	118	$\{a_n\}$
17	a/b		ϵ
20	$[a-b]$		N_ϵ
21	$\|c\|$	119	$\lim_{n\to\infty} a_n$
28	$\{L, R\}$		
36	$\aleph_0$	121	$\sum_{n=1}^{\infty} a_n$
41	(a, b)		
43	$n(z)$	124	$\lim_{x\to a^-} f(x)$
	$\|z\|$		$\lim_{x\to a^+} f(x)$
55	$R[k]$		
	$R(k)$	126	$\delta_{\epsilon c}$
57	$R^*(i)$	128	$p'(x)$
58	$b\|a$	129	$p^{(n)}(x)$
59	(a, b)	163	S_a
60	$[a, b]$	174	$[a_{ij}]$, $i, j = 1, 2, \ldots, n$
66	$b + a$	175	$\|a_{ij}\|$, $i, j = 1, 2, \ldots, n$
69	$\coprod$	176	P_n
76	334_6	178	(ab)
83	$a \equiv b \pmod m$	185	Σ
87	$a \not\equiv b \pmod m$	199	C_{n-r}^n
	$[r] \pmod m$	272	f_x

INDEX

Abelian group, 39
Absolute value, 21, 43
Acnode, 263
Addition, properties of, 10, 11
Adjoin, 54–55
Affine transformation, 224
Aleph zero, 36
Algebra, Fundamental Theorem of, 139
Algebraic complement, 199
  extension, 55
  function, 132, 264–269
  number, 26
  plane curve, 268
Algebraically closed, 49
Allowable coefficients, 102
Analytic functions, 133
Angle trisectors, 242–245
Approximate solutions, 169–171
Archimedes, angle trisection of, 242
  postulate of, 61, 109
Arithmetic, Fundamental Theorem of, 68
Arithmetic mean, 20
Associates, 103
Asymptotes, 264, 272
Augmented matrix, 208
Average, 20

Base, 25, 75–80
Binary operations, 5
  relations, 5
  system, 78–79
Binomial, 99
Bolzano-Weierstrass Theorem, 34
Bounded numbers, 33–34
Bounds for roots, 159, 163–164
Branch point, 269

Cantor-Dedekind Axiom, 32
Cantor's Theorem, 35
Cardan's formulas, 153
Cardinal numbers, 2–7
  transfinite, 35–38
Casting out nines, 85–86
Cauchy convergence criterion, 119–120
  sequence, 34, 119, 133
Chinese Remainder Theorem, 95
Circle, 254
  point, 254
  quadrature of, 239
Class of elements, 1
  of a permutation, 177
Classical constructions, 229–241
  assumptions for, 230
  basic, 233–235
Closed cut, 28
  interval, 115
  set under an operation, 5
Coefficients, 99
  set of allowable, 102
Cofactor, 188–189
Column expansion, 187
  index, 174
Commutative addition, 10
  group, 39
  multiplication, 13
Complementary minors, 199
Complex number, 41–42
  amplitude of, 46
  argument of, 46
  classification of, 53
  exponential representation of, 46
  imaginary part, 42
  quotient of, 43–44
  real part, 42
  trigonometric representation of, 46

Complex plane, 43
  n-space, 251
Component, 274
Composite number, 63
Composition of ordinates, 275
Conditional equation, 134
Cone, right circular, 255
Congruence, 83–95, 113
  class modulo m, 87–89
  linear, 93–95
  solvable linear, 94
Congruent modulo m, 83
Conic sections, 254–258
  degenerate, 255–256
Conjugate complex numbers, 43
  imaginary roots, 142–143
Constructions, 228–249
Continued fraction, 74–75
Continuous functions, 122–127
  at a point, 123
  graph of, 122
  on an interval, 124
  uniformly, 127
Continuum, cardinal number of, 38
Contradictory, 27
Contrapositive, 24
Contrary, 27
Convergence, Cauchy criterion for, 119–120
Convergent sequences, 119
  series, 121
Coordinates, 31–32, 46
Countably infinite sets, 36
Cramer's Rule, 172, 205–208
Cubic curves, 262–263
Cubic equations, 150–154
  reduced, 152
  resolvent, 155
Curve, algebraic plane, 268
  cubic, 262–263
  higher plane, 262–267
  quartic, 263
  reducible, 274
  unicursal, 263
Curve fitting, 277–280
Curve tracing, 246, 270–276
Cusp, 263
Cuspidal cubic, 263
Cuts, 28–30
Cyclic group, 52
Cylinder, 258
  right, 255

Decimal, exact, 25
  infinite nonperiodic, 26
  infinite periodic, 25, 82
  repeating, 25, 82
Decimal notation, 80–82
Dedekind cut, 28
  Postulate, 28
  Theorem, 28
Deficiency, 263
Defined operation, 5
Degree of a polynomial, 99
Delian problem, 239
De Moivre's Theorem, 50–54
Dense set, 19
Denumerably infinite set, 36
Derivative, 127–129
  partial, 272
Descartes' Rule of Signs, 156–160
Determinant, 174
  column expansion, 187
  evaluation, 195
  expansion, 192
  geometric applications, 217–227
  Laplace's expansion, 199
  notation, 174–175
  order, 185
  principal diagonal, 175
  row expansion, 183–185
Digits, 25
Dilation, 226
Dimension of Euclidean spaces, 250–251
  complex spaces, 251
Diophantine problems, 95–97
Directrix, 257

Discontinuity, 123–125
  finite, 125
  infinite, 125
  oscillating, 125
  removable, 124
Discriminant, 141
Divergent series, 121
Divides, 58, 102
Divisibility, tests for, 85–86, 135
Division, 16
  Algorithm, 60–63, 104–106
  sequence, 161
  synthetic, 135–137
Divisor, 12
  common, 59
  greatest common, 59, 109
Domain of a function, 115
Double points, 262–263
Duplication of cube, 239

Elementary symmetric polynomials, 141, 144–149
Elementary transformations, 204
Ellipse, 255–257
Elliptic paraboloid, 259
  cubic, 263
Empirical equation, 278
Equation, conditional, 134
  empirical, 278
  linear homogeneous, 205
  rectifying, 278
  systems of linear, 205–211
  theory of, 134–171
Equivalence relation, 7
Euclidean Algorithm, 71–75, 109–112
  n-spaces, 250
  transformation, 225
Euler's ϕ-function, 88
  Theorem, 92
Evaluation of a determinant, 195
Evolution, 16
Excluded region, 271
Exponential notation, 12

Factor, 12
  Theorem, 135
Fermat's Last Theorem, 95–97
  Simple Theorem, 92
Field, 39, 54–57
  adjunction to, 55
  algebraic extension, 55
  quotient, 55, 101
  skew, 39
Finite number, 23
  permutation, 179
Focus, 257
Folium of Descartes, 262–263
Function, 114
  algebraic, 132
  analytic, 133
  continuous, 122–127
  decreasing, 117
  discontinuous, 123–125
  domain of, 115
  graph of, 251–275
  increasing, 117
  inverse, 127
  multiple-valued, 115
  range of, 115
  single-valued, 115
  Sturm, 161
  symmetric, 149–150
  transcendental, 132

Geometric mean, 232, 234
  transformations, 221–227
Graphical solutions, 276–277
Graphs, 250–277
Greatest common divisor, 59, 70, 72, 103, 161
Group, 39, 52

Heine-Borel-Lebesgue Theorem, 34
Hessian, 273
Hindu-Arabic notation, 78
Homogeneous polynomials, 271
Homothetic transformations, 227
Horizontal asymptote, 264

Horner's method, 169–170
Hyperbola, 255–257
Hyperboloid of one sheet, 260
Hyperplane, 252

Ideal, 113–114
Identity element under an operation, 12
Identity relation, 102, 134
  transformation, 226
Imaginary numbers, 42
Incommensurable, 24
Independent polynomials, 103
Indeterminate, 98–102
Indicator of m, 88
Indirect proof, 24
Induction, principle of complete, 9
  mathematical, 9
Inequalities (*See* Order relations)
Infinite decimals, 25–26, 81–83
  sequences, 118
  series, 121, 131
  sets, 35–38
Initial, 99
Inner product, 201
Integers, 2, 22
Integral domain, 58
  points, 31
Intercept, 264
Interval, 115
Inverse elements, 15
  operations, 15–16
  transformation, 226
Inversion, 177
Involution, 16
Irrational numbers, 24, 26, 30
Irreducible polynomials, 106–109
Isoklinostat, 245
Isolated point, 263
Isolation of roots, 164
Isomorphism, 18

Kempe's angle trisector, 245
Klein's definition of a geometry, 221

Laplace's Expansion, 199
Least common multiple, 60, 70, 104
Least squares, method of, 278
Limit, 116, 118–122
Line of a matrix, 187
Linear birational transformation, 148
  combination, 195
  congruence, 93–96
  subspace, 252
Linear Equations, Fundamental Theorem for Systems of, 209
Linearly dependent, 213–214
  independent, 214
  ordered, 14
Linkages, 244–247

Mathematical induction, 9
Matrices, 173
  augmented, 208
  coefficient, 208
  cofactors of, 188–189
  equal, 223
  geometric applications of, 217–227
  lines of, 187
  minors of, 188, 198–204
  normal form of, 204
  notation for, 174
  order of square, 175
  product of, 201
  rank of, 202–203
  square, 174
  triangular, 197
Mean proportional, 234
Mersenne prime, 67
Minor(s), algebraic complement of, 199
  complementary, 199
  of an element, 188
  principal, 204
  rth, 198
Modulus of a complex number, 43
  of a congruence, 83
Monic polynomial, 103
Monomial, 98–99

Multiplication, properties of, 10–12
Multiplicity of roots, 164
  graphical intersections, 253

Nappes, 255
Natural number, 2
  order, 176
Negative numbers, 20–23
Newton Diagram, 272
Newton's method, 169
Nim, 79
Nodal cubics, 263
Node, 263
Nonnegative number, 19
Norm, 43
Normal form of a matrix, 204
Numbers
  absolute value of, 21, 43
  algebraic, 26
  bounded, 33–34
  cardinal, 2–7
  complex, 40–54
  composite, 63
  constructible, 236
  finite, 23
  imaginary, 42
  irrational, 24, 26, 30
  natural, 2
  numerical value of, 21
  ordinal, 2
  perfect, 60
  prime, 63–70
  pure imaginary, 42
  rational, 17–23, 81–82
  real, 24, 35
  signed, 21
  theory of, 58–97
  transcendental, 26
  transfinite cardinal, 35–39
  unbounded, 33
Number systems, 39–40, 54–57

One-to-one correspondence, 2
Open cut, 28
  interval, 115
Order-isomorphism, 18
Order of determinants, 185
  elements of cyclic groups, 52
  square matrices, 185
Order relations for complex numbers, 41
  nonnegative integers, 14–15, 17
  rational numbers, 17–24
  real numbers, 28–31
  transfinite cardinal numbers, 35
Ordered products, 225–224

Pantograph, 246–247
Parabola, 255–256
  semicubical, 263
Peano's postulates, 8–9
  space-filling arc, 275
Permanence, 157, 177
Permutation, 175
  class of a, 177
  even, 177
  finite, 179
  odd, 177
ϕ-function, 88–92
Plane figure, 235
Point reflection, 226
Polynomial equations, 134–171
  conditional, 134
  cubic, 150–154
  degree of, 139
  identity, 134
  isolation of roots, 163–164
  multiple roots, 164–168
  number of roots, 139–141, 156–164
  of degree greater than 4, 142
  quartic, 154–156
  rational roots, 148
  roots, 134
  solution of, 134, 139, 141–142, 166–171
  transformations of roots of, 146–150
Polynomials, 98–134, 144–150, 161, 252–264

Polynomials (*Cont.*):
associate, 103
derivative of, 128–129
elementary symmetric, 144–149
graphs of, 252–264
homogeneous, 188, 271
independent, 103
irreducible, 106, 143
monic, 103
reducible, 106
relatively prime, 103
ring of, 100
Sturm's, 161
symmetric, 149–150
zeros of, 134
Postulates for real numbers, 28–30
Prime numbers, 63–71
Primitive roots, 52, 87–89
solutions, 96–97
Principal ideal, 114
minor, 204
value, 46, 50
Product symbol, 69
of matrices, 201
Proof by elimination, 27
Pythagorean equation, 95–97
Theorem, 96

Quadratic equation, 151
Quadratrix, 239–240
Quadrature of circle, 239
Quadric surfaces, 255, 258–262
Quartic curves, 263
equation, 154–156
Quotient field, 55

Range, 115
Rank, 202–203
Rational functions, 101, 264–267
numbers, 17–23, 81–82
operations, 5
roots, 148
Real graph, 268
numbers, 26, 28
part of complex number, 42
Real polynomial equations, 159
Reciprocal modulo m, 94
Rectifying equations, 278
Reduced cubic equation, 152
Reducible curve, 274
polynomial, 106
Reductio ad absurdum, 24
Reflexive relation, 7
Relatively prime, 60, 88, 103
Remainder Theorem, 135
Residues modulo m, 87
class of, 87–89, 114
complete system of, 87
reduced system of, 88–89
Resolvent cubic equation, 155
Riemann surface, 268–269
Ring, 54–55
of integers, 58
of polynomials, 100
Roots, 134
bounds for, 159, 163–164
conjugate imaginary, 143
construction of, 236–237
isolation of, 163–164
multiplicity of, 164
of unity, 52, 87–89
simple, 164
transformations of, 146–150
Rotation, 221–222
Row expansion, 183–184
index, 174
Russian peasant multiplication, 78, 80

Secant, 128–129
Segment, 115
Sequence, 118
Cauchy, 119
convergent, 119
division, 161
limit of, 119
null, 118
Sturm, 161, 165–166

Series, infinite, 121, 131
Set(s), 1
closed, 5
continuous, 32
elements of, 1
empty, 4, 28
equivalent, 3
finite, 3
infinite, 3
mutually exclusive, 4
nonempty, 15
null, 4
proper subset of, 4
subsets of, 4
void, 28
well-ordered, 15, 61–62
Set-theoretic addition, 4
Sieve of Eratosthenes, 65
Simple root, 164
Singular point, 262
Slope, 128–129
Solutions, 17, 93, 96, 134
approximate, 169–171
graphical, 276–277
Spaces, complex, 251
Euclidean, 215, 250
four-, 260–261
Square matrix, 174
determinant of, 174
principal diagonal, 175
Straightedge, 229
Sturm functions, 161
polynomials, 161
sequences, 161, 165–166
Sturm's Theorem, 160–168
for multiple roots, 166
Subtraction, 16
Summation symbol, 185
Surfaces, quadric, 258
Riemann, 268–269
ruled, 260–261
Symmetric functions, 149–150
graphs, 271
polynomials, 149–150
Symmetric relations, 7
Synthetic division, 135–138, 159
Systems of linear equations, 205–211
augmented matrix of, 208
consistent, 207
determinant of coefficients, 205
Fundamental Theorem for, 209
matrix of coefficients, 208

Tangents, 128
at the origin, 271
Taylor's formula, 130
series, 130–133
Tomahawk, 244
Totient of m, 88
Transcendental function, 132
Transfinite cardinal number, 35
Transformations, geometric, 221–227
group of, 226
of roots, 146–150
ordered product of, 223–224
Transitive relation, 7
Translation, 221–222
Transpositions, 178–183
Triangular matrix, 197
Trinomial, 99
Trisection of any given angle, classical, 240–241
nonclassical, 242–245
Trisectrix of Maclaurin, 243

Unbounded numbers, 33
Unicursal curve, 263
Unique Factorization Theorem, 68
Units, 59, 63, 102
Unity, 12, 55
roots of, 52, 87–89

Vacuous graph, 251
Vandermonde's determinant, 198
Vector, 46
Variable, 93, 98
change of, 112–113, 138–139

Variable (*Cont.*):
continuous real, 115
dependent, 114
independent, 114
positive integral, 115
Variation, 157
Vertical asymptote, 264

Well-ordered, 15
Wilson's Theorem, 95

Zero, 12
cut, 29
deficiency, 263
divisor, 39, 59, 90
of a polynomial, 134